UNIVERSIDADE DE LISBOA
Faculdade de Ciências
Departamento de Informática

DETEÇÃO, IDENTIFICAÇÃO E CLASSIFICAÇÃO DE ATIVIDADE SUSPEITA EM BASES DE DADOS COM INFORMAÇÃO RESERVADA

Jorge Manuel Alves das Neves Lima

DISSERTAÇÃO

MESTRADO EM INFORMÁTICA

2014

UNIVERSIDADE DE LISBOA
Faculdade de Ciências
Departamento de Informática

DETEÇÃO, IDENTIFICAÇÃO E CLASSIFICAÇÃO DE ATIVIDADE SUSPEITA EM BASES DE DADOS COM INFORMAÇÃO RESERVADA

Jorge Manuel Alves das Neves Lima

DISSERTAÇÃO

MESTRADO EM INFORMÁTICA

Trabalho orientado pelo Prof. Doutor António Casimiro Costa
e co-orientado por Eng. José António dos Santos Alegria

2014

Agradecimentos

Devo agradecer ao Prof. António Casimiro e ao Eng. Alegria por se terem disponibilizado para orientar esta tese, bem como ao Pedro Inácio por me ter selecionado para este projeto. Tanto ele como o Bruno Possolo, perito residente em Guardium, foram instrumentais no seu apoio ao planeamento inicial do projeto.

Ao João Duarte, agradeço o acompanhamento diário que me providenciou durante a fase principal do desenvolvimento do projeto e também na elaboração da tese.

Sem dúvida, este projeto não teria sido possível sem o apoio incansável do João Seara, administrador de sistemas, e também do Bruno Miguel, administrador de bases de dados. Para ambos, um grande obrigado e louvor ao seu profissionalismo, à sua disponibilidade e capacidade de trabalho.

Na aprendizagem da linguagem Ruby, tenho necessariamente de incluir o Hugo Silva e o Tiago Sequeira nos meus agradecimentos pelas dicas que foram dando.

Ao Ricardo Ramalho e ao José Lourenço, devo agradecer pelo apoio à fase final de integração com o sistema Pulso para geração de visualizações.

A toda a equipa da PT, no geral, agradeço a amizade que me estenderam e o bom ambiente de trabalho proporcionado ao longo deste projeto.

Á minha avó.

Resumo

Vivemos numa época em que as novas tecnologias de telecomunicação estão cada vez mais no centro das nossas vidas. Simultaneamente, observam-se ataques aos sistemas de informação que armazenam os nossos dados e os transmitem.

Para além das ameaças externas, ataques por parte de pessoas estranhas ao sistema, existe a necessidade de detetar ameaças internas. Trata-se de fugas de informação confidencial ou sensível, possibilitadas pela necessidade de atribuir privilégios de leitura e escrita a quem administre bases de dados. Estas pessoas têm apenas tecnicamente, e não legalmente, o direito de aceder aos nossos dados, e esse direito deve cingir-se ao estritamente necessário para a manutenção e gestão dos sistemas na sua dimensão técnica, visando garantir a sua disponibilidade, desempenho, correção e conformidade a normas, regulamentos, ou leis.

O abuso de privilégios pode violar a privacidade de particulares, constituir espionagem industrial, ou até a divulgação de segredos de estado que comprometam a segurança de um país e as suas relações com parceiros internacionais ou aliados.

A não-conformidade com o contexto legal e regulatório sob o qual as empresas, através dos seus colaboradores, devem atuar, pode-lhes causar graves danos financeiros, manchar a sua reputação e responsabilizar criminalmente os seus executivos.

Neste trabalho, propomo-nos analisar métodos de deteção, identificação, e classificação de comportamentos suspeitos com base na tecnologia de monitorização de bases de dados *Guardium*.

É o objeto central deste trabalho o seu estudo e documentação num caso prático de uso: a monitorização de acessos a informação sensível alojada nas bases de dados da Portugal Telecom.

Para tal, pretende-se definir o comportamento correto de um administrador de bases de dados, e por contraste o que seja anómalo, levantando suspeitas de uso abusivo, focando principalmente os casos de consultas particularmente seletivas ou exportações massivas de dados.

Palavras-chave: segurança, bases de dados, monitorização, Guardium, aprendizagem

Abstract

We live in a time when new telecommunications technologies are increasingly at the center of our lives. Simultaneously, attacks are sometimes observed that jeopardize the systems that store and transmit our information.

Threats exist beyond the scope of external attacks on the system, by third parties outside the company. There is also a need to detect internal attacks, sensitive information leaks perpetrated by insiders who have access privileges to the databases they need to administrate.

Their right to access our data is only technical, not legal, and should be confined to the essential technical requirements of their systems maintenance and management duties, which are to guarantee availability, performance, correctness and conformity to the law, certain standards and regulations.

Abusing these privileges may violate people's privacy, constitute industrial espionage, or even compromise a country's national security and its relationship with its trade partners and allies.

A state of non-conformity with the legal and regulatory context under which corporations should operate, through the actions of their co-workers, may cause great financial damage, stain their reputation and hold their executives criminally responsible.

In this project, we propose to analyze methods to detect, identify and classify suspicious behavior based on the *Guardium* database activity monitoring software. It is the main goal of this project to study and document a practical use-case: monitoring access to sensitive information hosted by Portugal Telecom's databases.

To that effect, we intend to characterize the normal behavior for a user, and in contrast, to detect anomalies that may arise suspicions of abusive behavior, focusing primarily on particularly selective queries or massive data exports.

Keywords: security, databases, monitoring, Guardium, machine learning

Conteúdo

Índice de figuras

Índice de Quadros

Capítulo 1 Introdução

1.1 Motivação

A Portugal Telecom tem mais de cem bases de dados sob monitorização, usando duas soluções, chamadas Guardium e Imperva. Este trabalho foca-se no Guardium, que monitoriza acessos a dados sensíveis dos clientes dos serviços de telecomunicações, por parte de utilizadores nominais, privilegiados, ou aplicacionais.

É legítimo aceder a dados sensíveis para determinados fins e nenhum meio tecnológico pode determinar com exatidão as intenções de um utilizador. A maior parte do tráfego trata do uso e faturação das chamadas dos seus clientes, quer pela consulta do seu extrato mensal através de uma página de internet desenvolvida para esse fim, quer pela consulta facilitada por operadores de *call centers*.

A situação complica-se por haver aplicações que retransmitem o utilizador, autenticando-se com as suas credenciais, e simultaneamente haver aplicações que ocultam o utilizador original, usando um único utilizador para fazer consultas a dados sensíveis de variadíssimas pessoas e empresas, assumindo todos os comportamentos possíveis.

Segundo o dicionário da Priberam, a segurança é a qualidade do que está seguro, livre de riscos ou perigos, é o conjunto de ações que protegem algo ou alguém. O grupo de investigação ISECOM (*Institute for Security and Open Methodologies*) na versão 3 do seu *Open Source Security Testing Methodology Manual* (OSSTMM) define segurança como uma função da separação entre um bem (*asset*) e uma ameaça.

É difícil negar o interesse que o tema das ameaças de segurança tem tido a nível nacional e global. Portugal, *ex aequo* com o Reino Unido, foi o país que mais pesquisou pelo nome do grupo "Lulzsec" em Junho de 2011, de acordo com o Google Trends[1].

[1] https://www.google.com/trends/explore?q=lulzsec#q=lulzsec&cmpt=q

As quebras de segurança informática são hoje em dia, um assunto popular que transcende a própria ciência da computação. Só no jornal Público, encontramos mais de duzentos artigos relacionados com os Anonymous[2], grupo conhecido pela sua apologia da contestação social.

Estes grupos ficaram conhecidos pela divulgação de dados sensíveis, mas essas fugas foram tornadas possíveis através de ataques externos. Neste trabalho, pretendemos investigar fugas de dados que partem de ataques internos, feitos por pessoas da própria empresa que podem aceder rotineiramente ao sistema no seu dia de trabalho. Embora tenham tecnicamente as credenciais para aceder a estes dados, pode-se balizar a quantidade de dados que é legítima aceder, definir a forma como se acede, ou restringir a legitimidade da consulta apenas a acessos vindos de certos IPs através de certos protocolos.

Torna-se necessário, então, usar várias técnicas, ferramentas e soluções de segurança que atuam em vários aspetos de uma organização, em todos os pontos onde possa haver ameaças: monitorização e deteção de intrusões a nível da rede, configurações do sistema operativo, técnicas de desenvolvimento a nível da aplicação, e o registo de acessos a bases de dados.

Dentro destas categorias, a monitorização direta da atividade sobre bases de dados sensíveis, apoiada na tecnologia da IBM (Guardium), é o objeto central deste estudo.

Existem várias ameaças contra as quais a Portugal Telecom necessita de se proteger, como por exemplo:

- Fraude: Alterar a própria conta telefónica.
- Privacidade: Divulgar dados de clientes.
- Conformidade: Alterar informação sobre o negócio da PT que a comprometa perante o Ministério das Finanças.
- Segredo de Justiça: Divulgar quem está sob escuta, comunicações entre arguidos, ou a informação que foi escutada.

Examinemos em mais detalhe as várias dimensões sobre as quais a DCY atua.

[2]http://www.publico.pt/pesquisa?q=anonymous

1.1.1 Privacidade e dados sensíveis

A privacidade é um valor reconhecido no Artigo 12º da Declaração Universal dos Direitos Humanos: "*Ninguém sofrerá intromissões arbitrárias na sua vida privada, na sua família, no seu domicílio ou na sua correspondência, nem ataques à sua honra e reputação. Contra tais intromissões ou ataques toda a pessoa tem direito a protecção* [sic] *da lei.*"[3]

As violações da privacidade, através de fugas de informação sensível, podem ser feitas por pessoas com acesso privilegiado: administradores de bases de dados, de sistemas, ou programadores de aplicações.

Existe uma página precisamente para agregar fugas de informação, a Wikileaks, onde podemos encontrar resultados de fugas recentes e massivas: dois milhões de mensagens trocadas entre políticos da Síria[4].

O Guardium pode ajudar a garantir a privacidade, não só pela monitorização de atividade potencialmente suspeita sobre uma base de dados, mas também identificando e mascarando dados sensíveis. É possível definir padrões com expressões regulares, ou usar padrões pré-definidos. Assim, uma consulta a um número de telefone poderia devolver como resultado apenas asteriscos ou outro caráter usado como máscara.

Outra dimensão importante do problema da monitorização é garantir a integridade de dados que devem ser tratados de uma certa forma. Dados financeiros e dados clínicos têm de estar conforme o disposto na lei, ao abrigo de regulamentos internacionais que têm sido transpostos para a lei do país (Sarbanes-Oxley, HIPAA)[5].

1.1.2 Conformidade e auditoria

A atividade de uma empresa está sempre sujeita a um contexto legal. As empresas que lidam com dados sensíveis têm de ser visitadas regulares por parte de auditores independentes para verificar se a atividade da empresa está conforme o especificado nas

[3]http://dre.pt/comum/html/legis/dudh.html
[4]http://wikileaks.org/Syria-Files-PT-BR.html
[5]http://www.legalarchiver.org/

leis e normas relevantes. Chama-se "conformidade" à qualidade da situação da empresa em relação a esses compromissos jurídicos, ao objeto de avaliação do auditor.

A não-conformidade pode lesar gravemente as empresas às quais confiamos os nossos dados. Em 2012, a Anacom multou a Optimus em 6.6 milhões de euros por não cumprir uma norma[6] e a Comissão Nacional de Protecção de Dados multou-a em 7.5 milhões por uma fuga de informação[7].

Para facilitar a verificação desta conformidade, o Guardium disponibiliza ferramentas específicas de apoio à sua configuração, chamadas *compliance accelerators*[8] (aceleradores de conformidade).

Sendo que a conformidade tem de ser verificada regularmente por auditores, o Guardium também pode automatizar a geração de relatórios de conformidade, facilitando o seu trabalho.

É assim cada vez mais importante usar ferramentas como o Guardium, pelas poupanças que permitem ao automatizarem a auditoria, pela mitigação de riscos associados ao incumprimento legal, algo que acarreta consequências financeiras negativas.

1.2 Objetivos

Para cada base de dados (BD) de um painel pré-definido de bases de dados, com atividade sob monitorização da tecnologia *Guardium*, pretende-se investigar, implementar e documentar uma solução técnica que garanta à DCY a deteção, identificação e classificação de acessos que caiam em certos tipos de casos de uso.

É objetivo deste trabalho estabelecer um padrão de comportamento (*baseline)* para os utilizadores privilegiados de algumas BDs com dados reservados ou sensíveis. O Guardium já compila uma com base em métodos estatísticos (*Machine Learning*), mas há mais detalhes a ter em consideração, que o Guardium não analisa nem documenta em detalhe:

[6]http://www.publico.pt/economia/noticia/anacom-aplica-multa-de-66-milhoes-de-euros-a-optimus-1548820

[7] http://expresso.sapo.pt/caso-do-jornalista-espiado-optimus-pode-ser-multada-ate-75-milhoes-de-euros=f700216?_sm_au_=iHHsDZSnnvQP5RTR

[8]http://www.ibm.com/developerworks/data/library/techarticle/dm-1304pcidiss/

1 Acessos cujo volume de resultados com informação sensível seja superior a um determinado limite "normal", determinando anomalias em períodos homólogos.

Exemplo: Consultas massivas de dados sensíveis que levantem suspeitas de "extrusão" (fuga de dados), p.ex. listas completas de n^{os} chamados por alguém:

```
SELECT UNIQUE numtelefone FROM (…) WHERE numfactura=41111111
```

2 Acessos cujo SQL envolvido seja diferente do conjunto pré-existente de padrões.

Exemplo: Acessos extremamente específicos a dados sensíveis que podem indiciar fraude ou violação de privacidade, p. ex. nº chamado, nº chamador, hora, local:

```
SELECT chamado,chamador,antena FROM (…) WHERE msisdn=961111111
```

Para atingir estes objetivos, deverá ser desenvolvida uma solução em Ruby+MySQL que analise as queries SQL relevantes.

1.3 Entidade acolhedora - Portugal Telecom (DCY/AEC)

Este trabalho foi desenvolvido na Direção de Cybersecurity, Privacy e Business Continuity da PT Comunicações, no contexto da equipa de Arquitetura de Engenharia em Cibersegurança. A AEC é responsável pelo projeto DAMS (*Data Activity Monitoring and Security* – Segurança e Monitorização da Atividade de Dados) onde este trabalho se insere.

O projeto DAMS foca-se no uso de duas soluções de monitorização de bases de dados, o Guardium e o Imperva. Estas ferramentas para além de detetarem atividade suspeita também podem ser usadas na geração de relatórios para auditores. O comportamento suspeito é caracterizado pela resposta a cinco perguntas:

- Quem acedeu?
- Quando foi o acesso?
- O que foi acedido?
- De onde veio o acesso?
- Como foi feito?

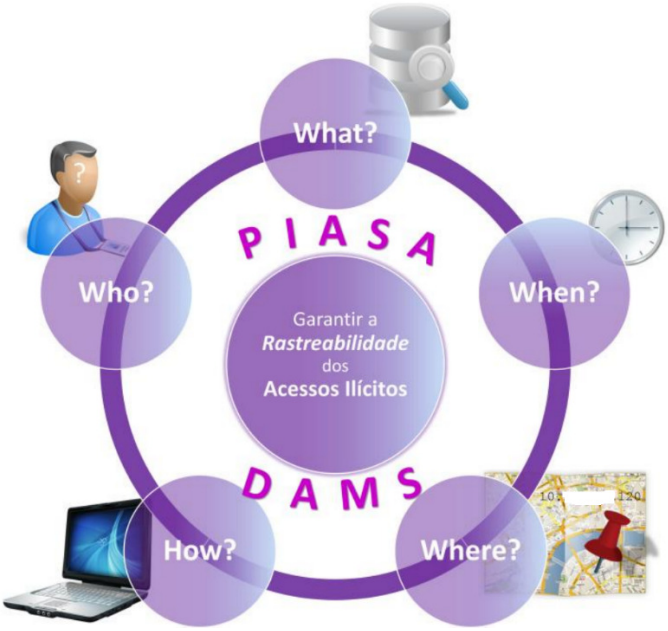

Figura 1 – *Projetos da DCY/AEC e sua caraterização do acesso a dados sensíveis*

A Figura 1 mostra como estas perguntas são respondidas pelo projeto DAMS em conjunto com outro projecto da AEC, o PIASA (PT *Identity & Access Security and Analytics* – Segurança e Analítica de Identidade e Acesso). Este outro projeto desenvolve

vários componentes para proteger a segurança física das instalações da PT, uso indevido de identidades (credenciais de acesso) pertencentes a terceiros, etc.

Outra equipa que faz parte da DCY/PT é a que desenvolve um sistema chamado Pulso [2] que recolhe métricas de desempenho dos vários sistemas da PT e gera diagramas para as visualizar de uma forma informativa. Embora não fizesse parte dos objetivos iniciais, foi desenvolvido um protótipo de integração entre o DAMS e o Pulso, comunicando dados sobre a quantidade de anomalias detetadas via HTTP para gerar "heatmaps".

1.4 Contribuições

Este projeto, desenvolvido de raiz e de uma forma autónoma, consiste numa plataforma baseada na web que permite visualizar listagens de comportamentos anómalos. Há um componente de extração e transformação de dados que corre diariamente, desencadeado pelo *cron*, e uma base de dados MySQL.

O valor que acrescenta ao esforço de investigação da Portugal Telecom é a oferta de funcionalidades que não estão presentes no Guardium, de uma forma extensível e interoperável, podendo de futuro abranger também a deteção de anomalias nos acessos monitorizados pelo Imperva.

Apesar de não estar diretamente integrado com outro software, depende dos *logs* (registos) do Guardium, e comunica estatísticas agregadas sobre os resultados para um sistema externo de visualização, o Pulso, via HTTP.

O projeto divide-se em dois subsistemas para dar resposta aos objetivos enunciados:

- O módulo **σ** (sigma) deteta anomalias no volume de dados retornado por cada acesso (*query*) à base de dados, compilando perfis de utilização para cada utilizador, em períodos homólogos.
- O módulo **κ** (kappa) sintetiza o texto integral das *queries* em padrões comuns, agregando-as em famílias que podem abranger milhões de queries cada uma, e deteta queries que não correspondem aos padrões mais usuais.

Mesmo não fazendo parte dos objetivos iniciais, foram também compiladas estatísticas agregadas para fazer *heatmaps*, ou "mapas de calor", permitindo visualizar rapidamente a evolução do sistema ao longo do tempo através da exibição de uma grelha que traduz a evolução do estado dos sistemas em cores, de acordo com a gravidade, ao longo do tempo.

Estes heatmaps residem no sistema Pulso e exibem contagens do número de anomalias observadas por dia, em cada um dos métodos utilizados:

- Volume de dados que exceda a soma da média com o triplo do desvio-padrão
- Nº de queries que fogem aos padrões previamente observados.

1.5 Organização do documento

Foi já exposta neste capítulo a motivação e objetivos do projeto, bem como uma descrição da equipa onde se insere, projetos da DCY com os quais interage e das nossas contribuições para eles.

No segundo capítulo, expõem-se trabalhos relacionados que contextualizam esta investigação e providenciaram enquadramento prático ao longo do desenvolvimento do projeto.

É feita uma introdução teórica no terceiro capítulo acerca dos métodos estatísticos usados neste trabalho e em áreas de conhecimento relacionadas, para melhor compreender as opções tomadas na sua resolução.

Todo o trabalho realizado é seguidamente enumerado no quarto capítulo, começando pelo levantamento da arquitetura de monitorização existente na DCY, e descrevendo em seguida as duas soluções desenvolvidas.

Antes de descrevermos cada uma das soluções, é feita uma contextualização dos problemas na sua dimensão técnica, e detalhada em abstrato a abordagem ao desenvolvimento das soluções.

Para finalizar, apresentam-se os resultados e tecem-se considerações acerca do valor do trabalho realizado.

Capítulo 2

Trabalho relacionado

2.1 Trabalhos anteriores desenvolvidos na DCY/PT

Para contextualizar este trabalho foi feita uma recolha das teses orientadas pelo Eng. Alegria em anos anteriores.

De grande relevo para este projecto, apresenta-se a tese de Francisco Ribeiro (2009) sobre descoberta e inferência de acessos anómalos a fontes de informação [1], com recurso ao motor de correlação Esper[9].

Contudo, no trabalho do Francisco Ribeiro (2009), a informação é tratada a baixo nível, fazendo monitorização direta do tráfego da rede (*Sniffing)*.

No caso do presente projeto, esse trabalho está muito facilitado pelo Guardium, que corre numa máquina separada (virtual ou física) chamada *Appliance.* Tem acesso a informação de rede através de um módulo chamado *Sniffer* mas também monitoriza diretamente o tráfego através de um módulo *kernel* instalado na própria máquina da base de dados, o que permite captar tráfego cliente-servidor quando o cliente corre na mesma máquina (Ex: aceder via SSH e lançar o cliente mysql no terminal).

Sendo este trabalho uma solução de monitorização, vamos examinar em seguida os vários tipos de soluções de monitorização mencionando alguns exemplos concretos de produtos existentes no mercado.

2.2 Soluções de monitorização

Existem muitos sistemas de monitorização e segurança com um leque muito diversificado de características. Alguns apenas monitorizam o desempenho e disponibilidade de um sistema, outros monitorizam o tráfego na rede, sendo que os mais completos podem intercetar e analisar SQL. Como se pode ver pela Figura 2, é mais comum investir na monitorização da rede do que na monitorização específica das bases

[9] http://www.espertech.com/

de dados, apesar de serem estas últimas as vítimas mais frequentes de ameaças. O projeto DAMS pretende responder a esta disparidade, investindo na deteção dessas ameaças.

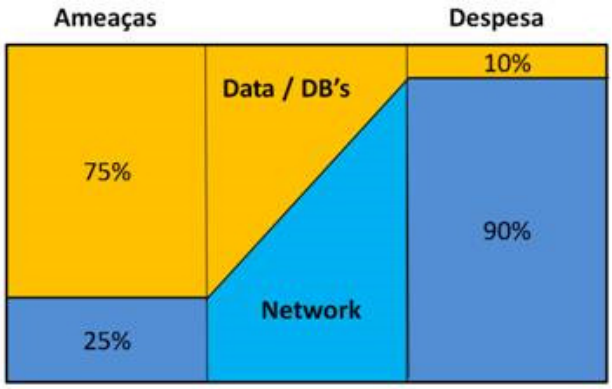

Figura 2 - *Desequilíbrio dos investimentos em segurança (fonte: apresentação interna da PT)*

Dentro destas soluções, independentemente de atuarem no domínio do tráfego em bruto que circula pela rede, ou serem específicas à atividade das bases de dados, existem vários tipos de programas com várias características:

- IDS (*Intrusion Detection System*), ou deteção de intrusão.
- IPS (*Intrusion Prevention System*), ou prevenção de intrusão.
- IDPS no caso de suprirem ambas as necessidades.

Estas soluções respondem a ataques externos. Contudo, um DBA não é um intruso, pode efetivamente aceder ao sistema, mas não deve aceder a informação sensível, muito menos alterá-la ou divulgá-la.

Há quem use outro designativo para sistemas que detetem este tipo de comportamento: MDS (*Misuse Detection System)*[3], porque apesar de o acesso ao sistema ser legítimo, o uso que se dá a ele pode não o ser.

No entanto, a terminologia MDS é pouco utilizada, visto que a sua funcionalidade é muitas vezes empacotada sob a égide de um IDS [4], ou usada para referir outros tipos de utilização ilegítima de computadores [5].

Tanto os sistemas de deteção de ataques externos como os internos podem ser baseados nos mesmos princípios teóricos de deteção de anomalias[4][5] que se pretende usar neste projeto.

O Guardium não só monitoriza bases de dados, como também as previne contra fugas de informação (DBLP - *Database Leak Prevention*). Utiliza-se no Guardium o

termo “extrusão”, por oposição à intrusão, para designar as atividades que conduzem a estas fugas. Vamos comparar com outra solução em uso na PT, o Imperva, e compreender a razão da escolha destes dois sistemas.

2.2.1 Monitorização de atividade em bases de dados

A DCY/PT utiliza duas grandes soluções de monitorização de atividade em bases de dados, o Guardium e o Imperva.

A compatibilidade dos produtos de segurança com várias plataformas é um requisito importante para estes ambientes heterogéneos, e apenas alguns são multiplataforma, sendo outros desenvolvidos pela mesma empresa que a base de dados, e fornecidos para trabalhar com a mesma marca.

A Forrester Research é uma empresa de investigação e assessoria que publica estudos dirigidos a diretores executivos. Em 2011 focou um dos seus estudos no sector de monitorização de atividade em bases de dados [26].

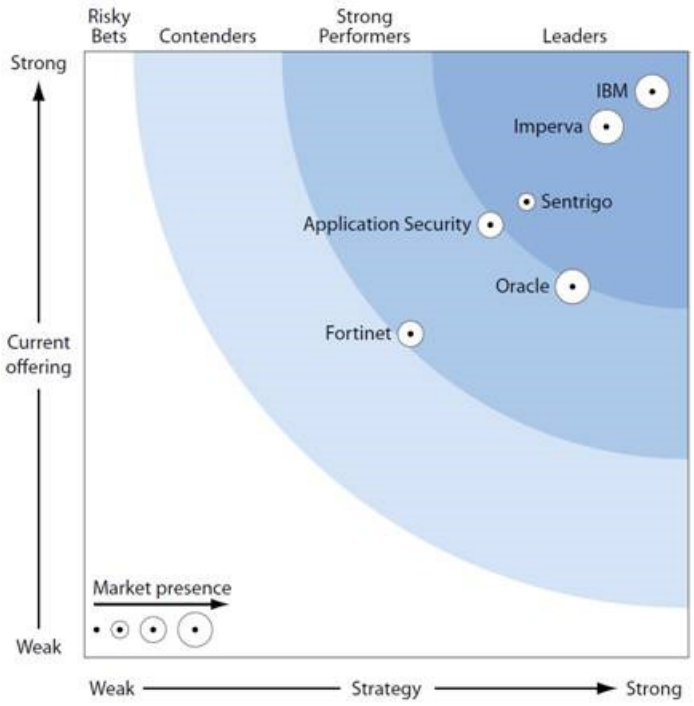

Figura 3 - *Qualidade da oferta e presença no mercado de várias soluções de monitorização [26]*

A Figura 3 sumariza os resultados do estudo desenvolvido, que menciona a existência de capacidades básicas de monitorização incluídas nas principais bases de dados. No entanto, a Forrester Research insiste em afirmar uma superioridade das soluções especializadas, segundo os seguintes critérios:

- Oferta presente: suporte para oito tipos de capacidades, como p.ex. auditoria das BDs, dos utilizadores, de aplicações, suporte a definição de políticas, repositório de auditorias, criação de relatórios e visualizações de análise estatística, proteção em tempo real, arquitetura, e facilidade de gestão.

- Estratégia: planos de expansão futura e estratégia de desenvolvimento tecnológico, bem como os recursos financeiros disponíveis para garantir a continuidade ao produto, o preço, e estratégia corporativa.

- Presença no mercado: a base de instalação (quantidade de clientes/sistemas a usar o produto), desempenho financeiro da empresa, serviços oferecidos, quantidade de colaboradores, parceiros tecnológicos, e presença internacional.

O Guardium foi considerado superior a nível da diversidade de capacidades tecnológicas de auditoria e gestão de políticas, embora o Imperva produza uma vasta riqueza de visualizações (*Analytics*) que não estão disponíveis no Guardium.

É essa uma das vantagens da nova versão do Guardium que é objeto deste trabalho. A nova versão do Guardium inclui um módulo de visualizações capaz de mostrar em tempo real o comportamento de um utilizador e avisar quando encontra comportamentos anómalos. Este projeto pretende atuar em conjunto com o Guardium e o Pulso para suplementar estas capacidades de deteção e visualização, aproximando-o do Imperva.

A DCY também utiliza soluções comerciais de monitorização de rede, bem como uma solução desenvolvida pela própria. Examinemos, por contraste, as suas funcionalidades.

2.2.2 Monitorização de rede

Sendo que um servidor de bases de dados é uma máquina como outra qualquer na rede, é possível usar ferramentas de monitorização de rede para monitorizar bases de

dados. Com efeito, o *Nagios*[10] esteve na base do sistema Pulso [2], e o *Splunk*[11] ainda está a ser usado na DCY. Mas, tal como em muitos outros mercados ainda em evolução, a segurança informática existe sobre um contínuo. Existem várias soluções com diversos tipos de funcionalidade. Graças à sua arquitetura extensível, o Splunk tem vindo a transcender a sua natureza inicial como solução de monitorização de rede, e incorporar funcionalidades específicas para bases de dados: monitorizar não só os padrões de acesso na rede, mas também o conteúdo das consultas[12] [13] [14].

Para lidar não só com o problema da segurança, mas também com o da privacidade, auditoria, conformidade com normas e legislação, é preciso adotar uma aproximação que use informação descritiva, como por exemplo:

- Indicar os campos de uma base de dados que são sensíveis.
- Diferenciar entre logins de sucesso, logins falhados, e acessos com consulta efetiva de informação sensível.
- Reportar a tabela e o campo acedido e quanta informação devolveu a *query* dessa consulta.

Uma simples análise estatística do tráfego não responde a estas questões. Para um programa de monitorização de rede, sem inspecionar aprofundadamente os pacotes de dados, enviar texto aleatório que não é interpretado como uma query é exatamente o mesmo que um acesso efetivo a uma base de dados. Não é possível, também, saber se o tráfego retornado pelo servidor é informação sensível, num sistema que contenha os dois tipos de dados.

Embora o Guardium também tenha características de monitorização de rede, torna-se clara, portanto, a vantagem de utilizar um programa específico para a monitorização de bases de dados, *versus* um programa de monitorização de rede, pelo que faz sentido ter um projeto e uma equipa dedicada a este tipo de tarefa.

De seguida vamos apresentar os conceitos teóricos por trás da realização deste trabalho, que também são usados em soluções comerciais.

---

[10] http://www.nagios.org
[11] http://www.splunk.com/
[12] http://apps.splunk.com/app/958/
[13] http://answers.splunk.com/answers/7065/splunk-for-sql-app-ie-database-activity-monitoring
[14] http://apps.splunk.com/app/874/

Capítulo 3 Conceitos

Neste capítulo iremos introduzir progressivamente alguns conceitos e métodos que podemos utilizar para caracterizar um sistema, traçando um percurso desde a análise estatística mais elementar, até à aprendizagem computacional, um ramo da inteligência artificial baseado em métodos de análise estatística mais avançados.

3.1 Análise Estatística

Para lidar com quantidades muito grandes de dados temos de construir agregados estatísticos, recolher métricas, encontrar padrões. Um comportamento anómalo define-se por contraste com um padrão de comportamento normal. Sendo que muitos fenómenos seguem padrões bem conhecidos, a normalidade pode ser modelada com base na premissa de corresponder a distribuições estatísticas familiares.

3.1.1 A distribuição normal (ou gaussiana)

Johann Carl Friedrich Gauss (1777-1855) descreveu[15] uma função de distribuição de probabilidades conhecida como "gaussiana", e que por estar na base de muitos fenómenos da natureza é conhecida também como "normal".

A abordagem mais simples à deteção de anomalias é precisamente assumir que a amostra tem uma distribuição normal, e calcular a medida na qual diverge da média. Como podemos ver pela Figura 4, numa distribuição normal podemos estimar que 95%, 99%, ou 99.9% dos valores estejam dentro de um certo intervalo delimitado pela soma da média com um múltiplo do desvio-padrão (representado usualmente pela letra grega *sigma*, minúscula - σ).

[15]http://books.google.pt/books/about/Theoria_motus_corporum_coelestium_in_sec.html?id=ORUOAAAAQAAJ

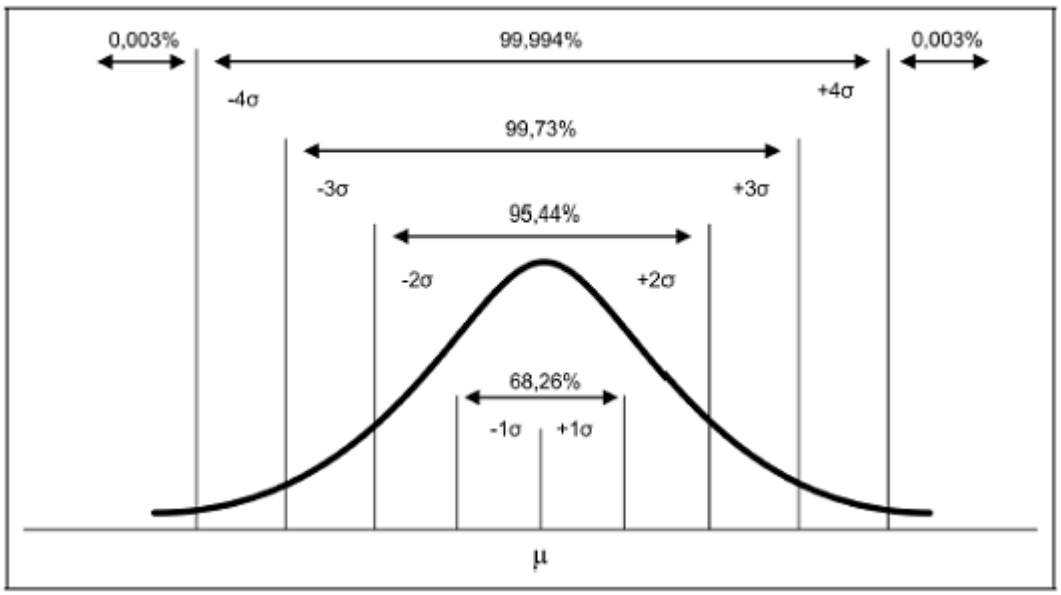

Figura 4 - – *Distribuição normal (ou gaussiana)*

A relação entre normalidade e anomalia é inteiramente subjetiva, depende do nível de qualidade que queremos garantir. Por exemplo, um bem de consumo fabricado à mão pode ter um número variável de defeitos, e ter como alvo aprovar 99% dos bens fabricados. Um sistema ou rede pode estar indisponível durante algum tempo, por falha, ou para manutenção planeada. Se garantirmos 99.9% de disponibilidade, isso equivale a quase 9 horas por ano de indisponibilidade. O grau de gravidade desta indisponibilidade é subjetivo: para um computador pessoal, pode ser aceitável, mas para um sistema de missão crítica não.

Como se pode ver pela Figura 4, é expectável que numa distribuição gaussiana haja uma determinada percentagem de valores a cair dentro de um raio em torno da média.

Assim, podemos considerar anómalos (*outliers* – também conhecidos como valores atípicos ou aberrantes) os valores que distam da média o dobro do desvio padrão (tolerando como normais 95.44% das amostras), ou o triplo (99.73% das amostras), pois não existe uma definição exata e universalmente aceite de anomalia [8].

Pode ser útil não modelar diretamente os valores, mas considerar antes que se obtém uma gaussiana a partir de uma transformação deles. Em particular, existe uma distribuição chamada log-normal que compara ordens de grandeza de valores. Examinemos em mais detalhe na secção seguinte.

3.1.2 Distribuição log-normal (ou de Galton)

A distribuição normal pertence a uma família mais alargada de distribuições chamadas elípticas. Uma distribuição importante a considerar é a log-normal, inicialmente proposta por Sir Francis Galton (1802-1911) a Sir Donald MacAlister (1854-1934), que a detalhou e publicou em 1879 [7].

É comum encontrar-se esta distribuição quando analisamos padrões de tráfego numa rede de computadores [6]. Com efeito, é esta a distribuição utilizada internamente para modelar acessos a bases de dados pelo Guardium, num algoritmo proprietário desenvolvido pela IBM Research [9] e assim faz sentido investigá-la.

Há boas razões para modelar a utilização nas bases de dados da PT como log-normal:

- Detetar fugas verdadeiramente massivas e anómalas de dados, e não extremos de variabilidade que podem ser normais.
- Tolerar a grande dispersão de valores que existem, e que abrange várias ordens de magnitude.

Diferentes tipos de clientes seguem diferentes padrões comportamentais. Há utilizadores casuais com muito poucas chamadas e há empresas com tráfego intenso. Podemos assim entender os clientes da PT como uma fatia transversal da sociedade que abrange empresas de vários tamanhos em várias fases de crescimento.

Segundo a lei da proporcionalidade de Gibrat (1931), a distribuição de empresas de acordo com o seu crescimento segue uma distribuição log-normal, proporcional ao presente tamanho da empresa, e independente do tamanho inicial [12]. Esta asserção foi validada empiricamente em Itália para os mercados de rádio, televisão, e telecomunicações [13], e foi assunto de uma recolha efetuada em 1997 por John Sutton [14], alargada a muitos mais casos.

Assim, existem muitas empresas pequenas e apenas algumas grandes, sendo a variabilidade da sua atividade algo que abrange várias ordens de magnitude.

Em 1948, C. E. Shannon, pai da Teoria da Informação, chamou *entropia* [10] ao grau de imprevisibilidade, ou aleatoriedade de um fenómeno, e em Estatística, a escolha de uma distribuição deve ser orientada pelo "*princípio da máxima entropia*", inicialmente

exposto por E.T. Jayne em 1957 [11]. A log-normal é a distribuição de entropia máxima dentro da sua classe.

O cálculo de indicadores estatísticos na distribuição log-normal processa-se da mesma maneira que na gaussiana: trata-se, afinal, de uma distribuição que se torna gaussiana após aplicarmos um logaritmo aos dados.

Caraterizar um sistema que evolui ao longo do tempo não se esgota em encontrar um padrão no espaço das frequências de dados. Temos de compreender também a sua relação com o tempo. Existem fenómenos com tendências lineares, semelhantes a funções lineares ($y = mx + b$) outros são cíclicos, semelhantes a funções trigonométricas ($y = sen(x)$), e também é possível serem uma mistura dos dois tipos ($y = m.sen(x) + b$). De seguida iremos focar-nos na sua caraterização.

3.1.3 Tendências e anomalias em séries temporais

Certos fenómenos evoluem linearmente ao longo do tempo, outros são cíclicos, ou compostos por uma combinação de ciclos e tendências lineares. Uma análise cuidada do passado pode expor estes padrões e ajudar a prever o futuro. Segundo a tese de Francisco Ribeiro (2009), é este o caso dos padrões de tráfego em rede, e é comum usar certos algoritmos, como o de Holt-Winters, com amortecimento exponencial [1]. Os algoritmos de deteção de anomalias apresentam características comuns:

- Capacidade de previsão dos valores seguintes na série.
- Quantificação do desvio entre a previsão e o valor real observado.
- Diferenciação entre desvios excessivos (anomalias) e desvios normais.
- Determinação do instante em que ocorreu a anomalia.

Muitas vezes, uma empresa conhece ou determina *a priori* a ciclicidade de um fenómeno, por estar associado ao normal decorrer da vida de trabalho de um empregado: um dia de trabalho que tem uma hora de início e fim definida, passando por uma pausa para almoço, a interrupção do ciclo diário de trabalho ao fim-de-semana, ou em feriados anuais. A previsão do futuro (*forecasting*) é importante para gestores, e o ramo da informática que lida com a recolha e visualização de indicadores chama-se Inteligência, podendo ser aplicada ao negócio na sua dimensão financeira, ás operações práticas da

empresa, e também à gestão da sua segurança, como uma das operações necessárias à continuidade do seu negócio. Vamos analisar a sua relação em mais detalhe.

3.2 Inteligência de Negócio, Operacional, e de Segurança

Sendo a estatística uma área instrumental, o uso dos métodos que foram expostos é importante em várias áreas, na intersecção das quais este projeto se encontra. O conhecimento por ela criado passa frequentemente pelas áreas de gestão de empresas, da operação de sistemas e redes, ou da sua segurança.

O termo *Business Intelligence* (BI) foi cunhado pela IBM em 1958 [17], num trabalho de Hans Peter Luhn (1896-1964), citando um dicionário para o seu uso da palavra "Inteligência": a capacidade de apreender factos apresentados de uma maneira que possa guiar uma ação em direção a um objetivo. Num meio empresarial em contínua evolução, os objetivos apoiados pela BI passam pela capacidade de comunicar efetivamente, inovar, liderar, ganhando adaptabilidade e pró-atividade face a novas situações[18]. Isto implica utilizar a Estatística não só de uma forma descritiva mas também para fazer previsões.

O típico processo de BI passa pela aquisição de dados gravados numa BD de factos chamada *Data Warehouse* para posterior processamento. A partir dessa BD constroem-se outros tipos de BD chamados *Data Marts* onde são agregadas estatísticas que foquem diferentes aspetos do negócio, e hoje em dia os sistemas baseados na web permitem acesso a essa informação de suporte à decisão a partir de qualquer sítio que tenha uma ligação à Internet [19].

A palavra "negócio" tem aqui uma vasta aplicação que abrange toda a atividade de uma empresa, não só na sua gestão comercial e financeira, como também na gestão das operações que a possibilitam. A aplicação de conceitos e metodologias de BI às operações do dia-a-dia de uma empresa, nomeadamente à gestão de redes, de sistemas de informação e da sua segurança é conhecida como *Operational Intelligence* [20].

As várias soluções de monitorização de sistemas e redes que atuam a nível de Operational Intelligence estabelecem ainda a diferença entre a mera monitorização de indicadores relativos ao desempenho, e os indicadores do "negócio" da Segurança. A esta área emergente chama-se *Security Intelligence*.

Exemplos de programa auto-descrito como tendo características de Security Intelligence incluem o Novell Sentinel[16], o SQL Stream[17], o Splunk[18] e o IBM QRadar[19], que tem possibilidades de integração com o Guardium.

Classificar um sistema como sendo de "Inteligência de Negócio" foca-o sob o ponto de vista da sua utilidade em providenciar indicadores de gestão. O mesmo sistema pode ser de "Prospeção de Dados" se enfatizarmos o aspeto mais técnico que lida com a sua recolha.

3.3 Prospeção de Dados (Data Mining)

A Estatística tem vindo desde sempre a enfrentar problemas que lhes são colocados por várias áreas científicas, tecnológicas e de negócio, como a Inteligência de Negócio.

Com o aparecimento e desenvolvimento dos Sistemas de Informação, a quantidade e complexidade de dados a processar sofreu um crescimento explosivo.

Os novos desafios de armazenamento, organização e pesquisa de dados que este crescimento colocou originaram o aparecimento de uma nova área de conhecimento chamada *Data Mining*[16] ou "Prospeção de Dados".

Por norma, *Data Mining* entende-se como o conjunto de seis tipos de tarefas[15]:

A Identificação de Anomalias ("*Anomaly detection*") é o cálculo, p.ex., de desvios em relação à média, outliers, ou outro tipo de valores fora do esperado que possam ser interessantes, constituir erros, ou de outra forma levantar suspeitas e motivar uma investigação adicional.

A Aprendizagem de Regras de Associação ("*Association rule learning*") é a pesquisa de relações entre variáveis. Uma loja online pode compilar perfis de utilizadores com os seus hábitos de consumo, determinando que tipos de produtos são comprados frequentemente em conjunto, e utilizar esta informação para efeitos de marketing.

O Agregamento ("*Clustering*") é a descoberta automática de grupos e estruturas nos dados, que são de alguma forma semelhantes, sem conhecimento prévio de qualquer lógica ou estrutura subjacente aos dados.

---

[16]https://secure-www.novell.com/products/sentinel/features/security-intelligence.html
[17] http://www.sqlstream.com/solutions/security/
[18] http://www.splunk.com/
[19] http://www-03.ibm.com/security/solution/intelligence-big-data/

A Classificação ("*Classification*") é a generalização de estruturas conhecidas aplicada a novos dados. Por exemplo, um cliente de correio eletrónico pode tentar classificar uma mensagem como "legítima" ou "*spam*".

A Regressão ("*Regression*") tenta encontrar uma função que descreva os dados com o mínimo possível de erro, ou desvio, em relação a eles.

A Sumarização ("*Summarization*") providencia uma representação mais compacta do conjunto de dados, incluindo visualizações gráficas e geração de relatórios.

Já vimos que o mesmo sistema pode enquadrar-se em duas áreas de conhecimento se focar a visualização de indicadores para efeitos de gestão, ou a mecânica da sua recolha. Falta ainda considerar os métodos de processamento que intermediam estes dois.

Caso um sistema de *Data Mining* faça uso de Inteligência Artificial, nomeadamente algoritmos de aprendizagem capazes de se adaptar a novas situações que emerjam ao longo do tempo, pode denominar-se "*Machine Learning",* ou aprendizagem computacional.

3.4 Machine Learning e Deteção de Anomalias no Guardium

As duas áreas de conhecimento chamadas *Data Mining* e *Machine Learning* têm muito em comum: os termos são por vezes usados como sinónimos e é também este o caso de uma publicação (ainda confidencial) que descreve o algoritmo do Guardium[9].

A IBM pretende afirmar o algoritmo de deteção de atividade maliciosa do Guardium como uma aproximação inovadora ao uso do *Machine Learning* para abordar o problema da atividade anómala no acesso a bases de dados.

Para tal, é usado um método com dois componentes:

1. Testar a consistência das ações de um utilizador com o seu passado.
2. Testar a consistência global com ações passadas de outros utilizadores.

Este motor de deteção é baseado num modelo probabilístico para estabelecer uma *baseline* de atividade normal. Para cada utilizador, o motor analisa as presentes ações e gera uma pontuação de anomalia com base no seu historial. Esta pontuação é baseada na raridade de cada ação, na sua intensidade, correlações entre ações diferentes, entre outras coisas.

Para aumentar o nível de precisão, é ainda utilizado um algoritmo de *Clustering* - o algoritmo *k-means* - para agregar utilizadores com base na semelhança dos seus comportamentos.

Desta forma, segundo um processo semelhante ao referido no capítulo anterior, a nova versão do Guardium efetua *Data Mining* aos factos em bruto e constrói visualizações, suplementando o *Data Warehouse* (armazém de dados) existente na versão anterior com um *Data Mart* (mercado de dados) orientado à deteção e visualização de anomalias em séries temporais.

Capítulo 4 O projeto Impervium

Neste capítulo vamos descrever o material informático e o fluxo dos dados através da rede que o integra, bem como a arquitetura geral do programa em termos da organização do armazenamento dos dados.

Numa segunda fase, vai ser descrita a abordagem a cada um dos objetivos inicialmente propostos, incluindo uma descrição mais aprofundada e técnica de cada um dos problemas que se põe.

Finalmente, apresentam-se as implementações dos sistemas desenvolvidos para resolver esses problemas.

4.1 Ambiente de monitorização dos sistemas

Para iniciar um contacto com o ambiente no qual ia ser desenvolvido o projeto, foi feita uma recolha acerca da arquitetura do sistema DAMS (Data Activity Monitoring and Security), que se apresenta aqui já integrado com os componentes desenvolvidos.

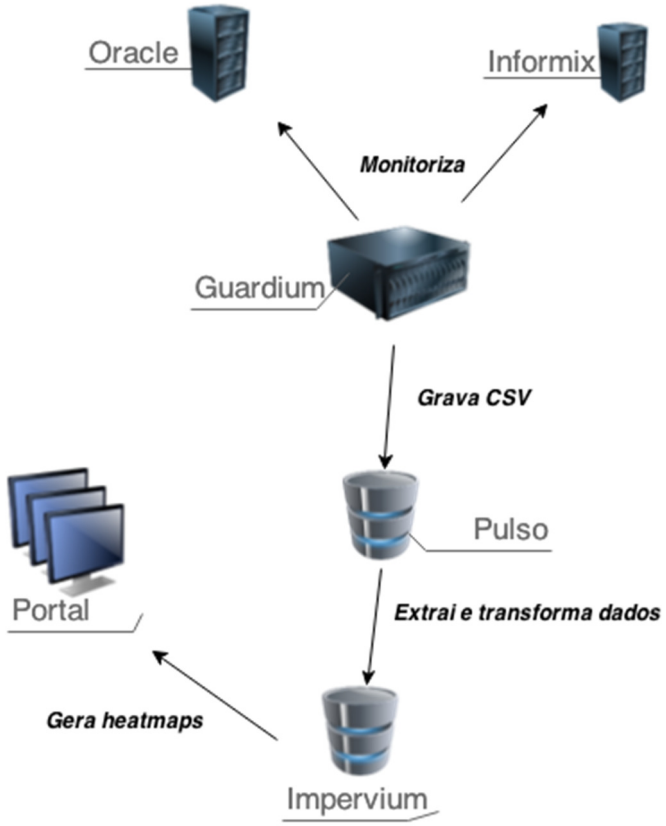

Figura 5 – *Integração dos sistemas Guardium, Pulso e Impervium*

Como se pode ver pela figura 5, o sistema de monitorização baseado em Guardium é composto por dois subsistemas. A monitorização é garantida por servidores Guardium autocontidos, que se dividem num agregador e dois coletores, um para cada tipo de bases de dados.

O agregador serve principalmente para integrar os dados dos coletores num só repositório, diariamente, de maneira a poderem ser gerados relatórios de toda a atividade da empresa. Pode ainda ser usado como ponto centralizado da administração do sistema de monitorização baseado em Guardium.

Os dois coletores recolhem dados, em tempo real, de dois tipos de bases de dados:

- Bases de dados IBM Informix sobre o sistema operativo HP-UX da Hewlett-Packard que representam mais de 90% do tráfego.
- Máquinas Oracle RAC (*Real Application Clusters*) que correm bases de dados Oracle sobre várias "zonas" (semelhantes a máquinas virtuais) do sistema operativo *Solaris*.

Nas bases de dados estão instalados módulos chamados GIM e S-TAP. GIM significa "Guardium Installation Manager" e é o que permite instalar ou gerir as S-TAP. A palavra "*tap*", em inglês, pode ser traduzida como "escuta", ou "sonda", sendo o componente que captura tráfego dentro da própria máquina, para o caso de haver aplicações instaladas na mesma máquina que a base de dados à qual acedem.

Esta "escuta de segurança" é composta por três módulos:

- PCAP: "packet capture", captura pacotes de dados que circulem pela rede.
- K-TAP: "kernel tap", captura tráfego local ao servidor, para o caso de haver uma base de dados no mesmo servidor que o programa que lhe faz consultas.
- A-TAP: "application tap", efectua monitorização integrada com a base de dados e específica a ela.

Em cada máquina, física ou virtual, existem IPs de acesso público e IPs que dão acesso ao interface com privilégios de gestão. Esta divisão de papéis em redes separadas é uma medida de segurança que impede qualquer tráfego público de entrar na rede de administração e vice-versa.

A monitorização funciona em regime de "dupla venda": É importante salientar que as bases de dados monitorizadas não são da DCY, nem é responsabilidade da DCY desenvolver os programas que correm sobre elas. A DCY não tem acesso ao código-fonte das aplicações nem ao esquema geral das bases de dados, tabelas, e campos. Apenas são conhecidos os que têm dados sensíveis.

Salvo uma ou outra exceção em que é preciso rastrear uma anomalia, a DCY não sabe quem são os indivíduos que acedem a estas bases de dados. Com efeito, parte do sistema de monitorização cifra os nomes de utilizador com que os colaboradores se autenticam. A investigação de potenciais acessos ilegítimos sai fora deste projeto, sendo conduzida a níveis superiores de gestão.

4.2 Política de segurança

O Guardium recolhe vários tipos de dados, desde métricas acerca do seu próprio funcionamento, consultas a dados sensíveis, passando pelas meras tentativas de entrada no sistema (*login*) quer tenham sucedido ou falhado. Relativamente ao comportamento do sistema, são monitorizados os sete indicadores apresentados no Quadro 1:

Tipo	Descrição
Mysql Is Up	Verifica se o MySQL está no ar, e devolve 0 se assim for.
Mysql Disk Usage	Percentagem de espaço em disco usado pelo MySQL.
Logger Queue Length	Tamanho da fila de espera do Logger: quanto menor, melhor, de preferência 0.
Sniffer Packets Dropped	Pacotes descartados pelo Sniffer.
% CPU Sniffer	Percentagem de tempo de processador utilizada pelo Sniffer.
% CPU Mysql	Percentagem de tempo de processador utilizada pelo MySQL.

Quadro 1 - *Indicadores das Appliances Guardium*

Segundo o manual do Guardium, é recomendável que a taxa de ocupação de disco do MySQL ronde, no máximo, 50% sem nunca exceder 90%. De outra maneira pode haver problemas de estrangulamento na BD.

São também recolhidos indicadores semelhantes para o *Sniffer* (sonda que captura pacotes da rede), percentagem de tempo de processador, memória ocupada, com especial

atenção para a sua capacidade de comunicação por rede. Não convém que quaisquer pacotes do Sniffer sejam perdidos. Assim, o Guardium mantém registo de pacotes que tenham sido ignorados ou moderados ("*Throttled*").

É ainda monitorizado o tamanho da fila de espera do *Logger*, o componente que efectua os registos. Idealmente será tão curta quanto possível, números elevados podem indicar sobrecarga dos agentes para o coletor.

Finalmente, são mantidos registos dos endereços IP dos agentes, o seu estado de atividade (up/down), e quaisquer erros do sistema S-TAP, expostos no Quadro 2:

Evento	Descrição
Event Type	Sucesso, tipo de erro, etc.
Event Description	Descrição do evento
Timestamp	Data e hora em que o evento ocorreu

Quadro 2 - *Eventos S-TAP (Agente Local)*

Os eventos mais graves são os sinalizados com "DISCONNECT/CONNECT". Se não forem provocados por uma reinicialização voluntária do agente, poderá haver perda de eventos enquanto o agente não acabar de reiniciar.

O Guardium monitoriza acessos a dados sensíveis segundo uma política de segurança (*security policy*) - um conjunto de regras (*security rules*) a aplicar ao tráfego entre as bases de dados e seus clientes, ferramentas de gestão ou aplicações em geral. As regras podem ser aplicadas condicionalmente aos pedidos que são feitos à BD ou a respostas devolvidas pelas mesmas. A condição aplicada pode ser uma simples verificação de endereço IP, para averiguar se pertence a um conjunto de utilizadores autorizados, ou algo muito mais complexo. Outros tipos de dados a verificar incluem p. ex. o nome do utilizador ou aplicação que está a tentar aceder à BD, o tipo de pedido efetuado, ou mesmo a hora do dia.

O Quadro 3 sumariza a política de segurança Guardium utilizada na DCY:

Campo	Descrição
Rule Position	A ordem em que vai ser verificada a regra.
Rule Type	As regras podem ser do tipo Access, Exception, ou Extrusion: • Uma regra de acesso aplica - se a pedidos de clientes (p. ex. um comando SELECT ou UPDATE). • Uma regra de exceção é aplicável a respostas do servidor (p. ex. falhas repetidas no estabelecimento de uma ligação). • Uma regra de extrusão atua sobre os dados retornados pelo servidor, como p. ex. padrões alfanuméricos.
Policy Description	Nome descritivo da política.
Rule Description	O nome da regra, um descritivo, para ser facilmente entendida por humanos. O próprio sistema pode sugerir nomes (tal como pode sugerir regras inteiras).
Client IP / Group	Um endereço e máscara de rede ao qual a regra se deve aplicar. Opcionalmente pode ser indicado um "Not" na configuração da regra, para aplicar a regra a todos os endereços menos esse.
Command / Group	O comando ao qual o pedido deve corresponder. No nosso caso, "SELECT".
Object/Field Group	Para a regra ser verdadeira, o objeto deve ser membro de um grupo especificado.
Service Name / Group	Serviço ou grupo de serviços ao qual o pedido ou resposta diz respeito.
DB Name / Group	O nome da base de dados ou grupo de BDs ao qual a regra se aplica.
Pattern/XML Pattern	Uma expressão regular que define um padrão de dados a identificar. Pode ser construída com o apoio de uma ferramenta interativa ou introduzida manualmente.
Action	Ação a tomar quando a regra é avaliada como verdadeira.
Rec. Vals	Indica se os dados que desencadearam a ativação da regra devem ou não ser registados.

Quadro 3 - *Política de segurança implementada em Guardium*

Numa fase inicial, foi desenvolvida uma aplicação para visualizar as métricas de desempenho previamente apresentadas, cujo objetivo foi apenas ganhar um primeiro contacto com a linguagem Ruby e com o ambiente de monitorização.

No entanto, o principal objetivo deste trabalho é tratar de casos de violação da política de segurança. É a partir de eventos que violem a política de segurança que o registo de dados é desencadeado – o sistema não monitoriza a totalidade do uso de uma base de dados, apenas o acesso a "objetos" configurados como sendo sensíveis, sejam tabelas ou campos de dados individuais.

Assim, a máquina Guardium regista os seus dados numa base de dados interna que guarda cópias dos seus registos no Pulso via SSH de 15 em 15 minutos.

Foi detalhado o ambiente Guardium com o qual este projeto se integra. Vamos fazer de seguida uma breve análise dos problemas a resolver por tratamento dos dados que o Guardium gera, e as abordagens tomadas à sua solução.

4.3 Análise e Desenho

O objetivo geral deste projeto é a identificação de comportamentos suspeitos, ao longo de duas principais dimensões, que se podem caraterizar *grosso modo* como extrações de dados anormalmente grandes (*dumps massivos*), ou consultas a dados sensíveis com queries anormalmente raras. Examinemos em mais detalhe cada um destes problemas, bem como o percurso experimental que nos levou à solução final efectivamente desenvolvida e posta em produção.

4.3.1 Identificação de dumps massivos

No primeiro caso de comportamento suspeito, os "dumps massivos", a abordagem utilizada foi a elaboração de estatísticas sobre o tamanho usual das consultas por cada utilizador, num dado período de tempo.

Foram utilizados os vários métodos descritos na introdução: modelando com distribuição normal/gaussiana, bem como log-normal, quer globalmente quer alinhando os dados segundo períodos homólogos.

Construíram-se, portanto, duas bases comportamentais de acordo com duas distribuições estatísticas, sendo os resultados examinados com vários níveis de granularidade.

Calcularam-se anomalias para ambas as distribuições, com vários níveis de tolerância: conforme mencionado em capítulos anteriores, não existe uma definição única e universal de anomalia, existem graus de anomalia, de raridade, da qualidade de serviço que se quer oferecer. São usados vários múltiplos do desvio-padrão para tipificar a distância à média (ou seja, à normalidade).

Foi construído um site para suportar as operações do dia-a-dia, permitindo a geração de relatórios segundo os vários métodos estatísticos e granularidades, divididos por tipo de utilizador.

4.3.2 Identificação de queries suspeitas

É objetivo deste trabalho a identificação de queries anómalas, que não pertençam a determinados padrões pré-estabelecidos. Nomeadamente queries muito específicas que procurem obter ilicitamente dados de um utilizador, ou que constituam dumps massivos.

A solução inicialmente proposta tentava encaixar o processamento deste texto em métodos estatísticos usados pelo Guardium para deteção de anomalias (*k-means clustering*) usando para comparação entre queries o algoritmo NCD (normalized compression distance) que poderia detetar queries muito diferentes da norma.

O algoritmo NCD compara dados binários resultantes da compressão do texto com a biblioteca Zlib. A comparação de strings depende do comprimento (strings de comprimentos diferentes são necessariamente diferentes) e há uma vasta gama de comprimentos desde queries na ordem dos 100 bytes até 30 KB.

Foi processada a maior família de queries (1.87 milhões), obtendo um "worst-case scenario", ou seja, considerando o pior dos casos possíveis em termos do desempenho exigido do sistema. Escolhe-se como termo de comparação a primeira query do ficheiro, e comparam-se todas as seguintes a essa.

Com este conjunto o NCD demorava 3h47m a processar, e verificou-se que dentro da mesma família de queries podia haver uma diferença de cerca de 50%. Ou seja, dentro de uma família de queries com o mesmo comprimento, e correspondentes ao mesmo comportamento de consulta de dados, metade dos caracteres poderiam ser semelhantes, e metade poderiam ser diferentes. Esta grande variabilidade não permitiria estabelecer um padrão do que é normal e distinguir o que é anómalo.

Usando o comando *diff* foi feita uma comparação entre a query pivot e a mais distante. A diferença entre as queries consistia inteiramente de algarismos, como datas, telefones, entre outros, sendo que todo o conjunto de números podia ocupar efetivamente 50% da query sem no entanto adicionar qualquer informação sobre o seu comportamento.

Foram substituídos os algarismos por *underscore* com o comando "sed s/[0-9]/_/g", um *wildcard* SQL que permite comparar queries com a cláusula LIKE.

Este filtro reduziu a amostra de dados a um único padrão de query, em menos de um segundo, e aplicando aos dados todos obtemos cerca de 1500 padrões. No entanto a comparação de queries ainda era lenta, pelo que se armazenaram apenas os *checksums* CRC32 (um inteiro 32 bits), técnica usada também por anti-vírus.

Foi assim escolhida uma aproximação baseada em assinaturas (signature-based detection) e não anomalias estatísticas baseadas em médias de distâncias entre strings.

4.3.3 Armazenamento de dados

A figura 6 tenta dar uma ideia conceptual de alto nível acerca do esquema geral de ETL (extração, transformação, carregamento) e armazenamento de dados. Trata-se de uma abstração construída em cima de um possível diagrama de classes, agrupando algumas em pacotes.

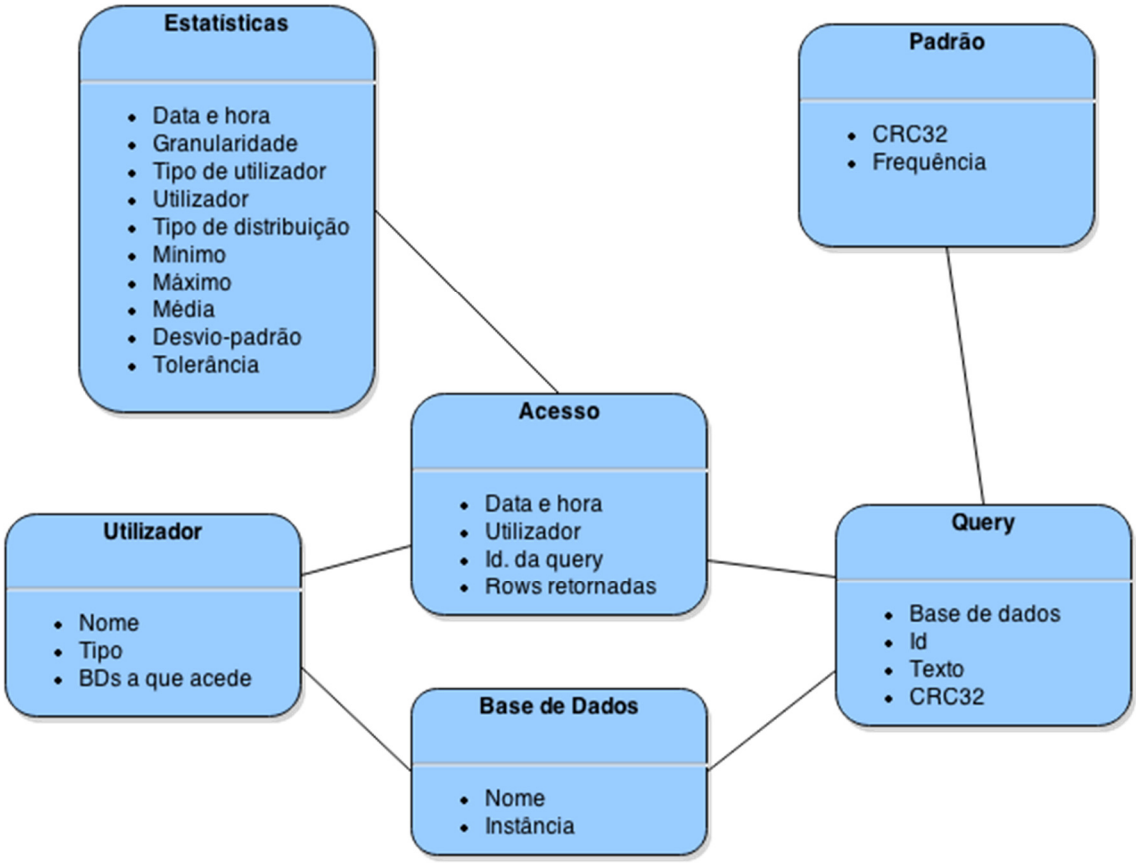

Figura 6 –*Dados extraídos dos CSV do Guardium e armazenados em MySQL*

Para maior facilidade de compreensão, a figura 6 não tem correspondência direta com tabelas (as estatísticas, acessos, e utilizadores estão divididas em várias tabelas por questões de optimização e tolerância a faltas) nem corresponde ás classes desenvolvidas em Ruby. Boa parte do projeto consiste de shell scripts e queries SQL, pelo que não é orientado a objetos, sendo que o Ruby foca-se na geração do interface HTML. Este tipo de tarefa de processamento de texto e cálculos estatísticos é muito mais fácil de

implementar e computacionalmente rápida na execução, ao usarmos as ferramentas UNIX e as funções estatísticas que fazem parte da linguagem SQL.

A entidade central dos dois sistemas desenvolvidos é o conceito de acesso: um utilizador faz um acesso a um dado instante temporal, através de uma query com um identificador numérico, obtendo um dado tamanho de resultados. Do lado esquerdo da figura 6, em cima, temos um pacote de estatísticas que recolhe vários indicadores com vários níveis de granularidade: médias globais, diárias, horárias, e em períodos de 10 minutos. Em baixo, temos metadados sobre os utilizadores e as bases de dados existentes, para efeitos de rastreabilidade ou para categorizar estatísticas agregadas por BD ou tipo de utilizador.

Do lado direito, de uma forma abreviada, temos o sistema de deteção de queries raras, ou anómalas. Mantemos um cache do texto integral das queries do dia anterior, compilamos padrões com checksum CRC32, contabilizamos a frequência de uso de cada padrão e associamos as queries aos acessos e ás bases de dados a que dizem respeito.

Vamos analizar de seguida os sistemas construídos sobre este modelo de dados em mais detalhe.

4.4 Implementação: Deteção de *dumps* massivos

Existe a possibilidade de utilizadores fazerem legitimamente dumps massivos, p.ex., fazer um backup de milhões de rows todos os dias à meia noite.

Compilando médias e desvios-padrões para períodos homólogos, ou seja, tirando médias separadamente sobre a quantidade de dados consultados a cada hora, podemos identificar estas situações.

Por comparação, foram feitas também médias globais, diárias, e para períodos de 10 minutos, ainda mais específico e sensível do que para uma hora.

Assim, para estudar as várias possibilidades, no interface do Impervium podem ser seleccionados como anomalias os acessos que excedam dois, três ou quatro desvios-padrões.

Para os anomalias na dimensão dos acessos, é oferecida uma classificação com um grau de gravidade correspondente ao número de desvios-padrões que o acesso ultrapassou (inspirados no conceito estatístico de intervalo de tolerância) e é adicionalmente identificada a query que o desencadeou.

Como objetivo secundário, pretende-se agregar a evolução destes comportamentos em indicadores gerais a representar graficamente por *heatmaps*, divididos por tipo de utilizador e base de dados, para serem usados pelas camadas de gestão.

Para tal, são extraídos dados a partir de várias fontes, nomeadamente vários tipos de logs do Guardium (formato CSV), ou de uma base de dados central que agrega dados do Guardium, do Imperva, e acrescenta alguns meta-dados que os caracterizam.

Dado o enorme tamanho destes dados, e algumas redundâncias que lhes são inerentes, torna-se necessário um forte investimento na sua filtragem.

4.4.1 Importação de dados

O Guardium está configurado para depositar vários tipos de logs numa directoria /home/sec/guardium em formato CSV, sendo esta a fonte principal de dados. Examinemos os seus esquemas de nomes e uma descrição dos seus conteúdos:

- G_URIGHTS_[INFORMIX|ORACLE]_[D|S|N] – estes logs são actualizados diariamente com a listagem de utilizadores por tipo de bases de dados e por tipo de utilizador: D para os DBAs, S para os "super users", e N para os "normais", ou seja, sem privilégios especiais atribuídos.

- G_CMD_OBJ_SENSITIVE – este tipo de ficheiro regista todos os acessos, rows retornadas, identificador numérico da query utilizada, sendo gerado um novo a cada 15 minutos. Podem atingir centenas de megas em alturas de tráfego intenso.

- Base de dados central – existe uma base de dados na máquina "Pulso" com parte destes dados, aos quais agrega ainda dados do outro sistema de monitorização, o Imperva, bem como uma série de metadata. Esta máquina serve de backend para o Pulso.

A espinha dorsal desde projecto é o MySQL. À semelhança da base de dados central, bem como de uma base de dados interna da appliance Guardium, foi instalada uma BD MySQL num servidor dedicado a este projecto. É utilizada para várias funções, armazenamento dos dados em bruto ou agregado, algum trabalho de parsing, operações de estatística.

A linguagem Ruby não foi planeada para processar grandes conjuntos de dados com garantias de desempenho, pelo que se investiu fortemente no uso das ferramentas de processamento de texto UNIX, e em queries SQL.

Os logs armazenam o mesmo acesso uma vez por cada campo sensível. As queries chegam a ultrapassar 30 Kbytes, pelo que estas repetições podem distorcer a distribuição estatística dos acessos ao repeti-los um número indeterminável de vezes, o que adulterava a distribuição estatística, contabilizando mais acessos do que efetivamente existiam.

Assim, optou-se por tipificar o comportamento eliminando repetições. A partir dos ficheiros G_CMD, selecciona-se apenas o timestamp, o nome do utilizador, o nº de rows retornadas, e o id da query. Esta selecção é feita com um simples comando "awk".

No caso do primeiro dia registado, isto reduziu 9 GB de ficheiros a apenas 17 MB.

Os ficheiros G_URIGHTS são importados directamente com um simples mysqlimport, diariamente.

4.4.2 Classificação de utilizadores

Há cerca de 6 milhões de acessos por dia, pelo que a tabela era extremamente lenta e as estatísticas demoravam demasiado tempo a calcular para serem geradas em tempo real.

O tamanho dos dados implicou uma preocupação adicional com desempenho. Inicialmente, só haviam dois tipos de utilizadores, conforme mencionado nos objectivos: DBAs e "power users" / "super users".

Para optimizar a tabela *sql_dams* foram criadas duas tabelas para cada um dos tipos de utilizadores (*d_sql_dams* para os DBA e *s_sql_dams* para os super users), ficando ambas as tabelas com cerca de 50% dos dados.

Os ficheiros G_URIGHTS identificam o tipo de utilizador, e são carregados diariamente para uma tabela. Estão sempre actualizados, pelo que a tabela é recriada de raíz todos os dias. Os dados são classificados fazendo um JOIN da tabela principal, sql_dams, com a tabela g_urights respectiva, *g_urights_d* ou *g_urights_s.*

A maior parte destes utilizadores eram aplicações configuradas com privilegios desnecessários, pelo que foi sugerido tratá-las como um caso à parte. Como não vêm identificadas nos ficheiros G_URIGHTS, foi criada uma whitelist manualmente. As

aplicações têm geralmente nomes conhecidos, e quando não são familiares, são óbvios. Os nomes de utilizador seguem todos um esquema XYnnnnnn, ou seja, duas letras seguidas de seis números, ao passo que as aplicações têm um nome descritivo facilmente identificável.

Assim, por coerência, foi criada uma whitelist *g_urights_informix_apps* e uma tabela *apps_sql_dams*. Isto tornou as tabelas anteriores obsoletas: existem muito poucas consultas de DBAs e super users. Após dois meses, apenas foram identificados 7 SELECTs de DBAs e 70 de super users, em milhões de acessos, os suficientes para serem examinados manualmente, sem terem devolvido um número de rows particularmente grande nem terem usado uma query particularmente suspeita.

Não sendo os acessos aplicacionais particularmente preocupantes, foi ainda tomada em conta a análise de utilizadores "normais", detalhados no ficheiro G_URIGHTS_INFORMIX_N e importado para uma tabela semelhante, *g_urights_informix_n*. Estes utilizadores têm acesso a dados sensíveis porque há tabelas configuradas como sendo de acesso público. No entanto, os acessos foram também considerados pouco preocupantes, visto termos obtido informação de que não são os próprios utilizadores a aceder, existindo simplesmente uma aplicação que faz os acessos em seu nome, em vez de utilizar uma conta especial só para a aplicação.

Os dados dos ficheiros G_SQL também eram classificados de uma forma semelhante, antes de terem sido abandonados.

4.4.3 Transformações e cálculos em SQL

Após carregamento para a BD do Impervium, são eliminados da tabela *sql_dams*, ainda antes da classificação, acessos que não devolvem dados (rows retornadas <= 0), com uma query do tipo DELETE FROM - este campo pode ser preenchido pelo Guardium com um número negativo no caso de haver erros, ou de serem executadas queries diferentes de SELECT que nunca poderiam devolver dados, por exemplo CREATE TABLE.

A partir dos acessos, foram calculadas estatísticas sobre os vários tipos de utilizador, com vários níveis de sensibilidade (dois, três, ou quatro desvios-padrões) e as duas distribuições atrás referidas (normal e log-normal). Foram também calculadas quatro granularidades, estatísticas globais para todos os acessos, estatísticas diárias para visualizar uma baseline com evolução de dia para dia. Traçaram-se também estatísticas

de períodos homólogos, para períodos horários, e para períodos de 10 minutos – foi constatado que existe uma enorme variabilidade diária, conforme se pôde ver pela análise exploratória.

As comparações homólogas muitas vezes não têm objectivo de fazer cálculos precisos, em Business Intelligence são muitas vezes usadas simplesmente para mostrar evolução de um indicador de uma semana para outra, com um pequeno número de barras no gráfico e uma linha de tendência. No entanto, os períodos homólogos aqui são calculados para todas as amostras no mesmo período desde o início. Ou seja, no caso das estatísticas horárias todos os acessos que ocorreram entre as 00h00 e as 01h00, por exemplo, compilando-se médias transversais a todos os dias desde o início do projecto.

Quer o parsing quer as operações de cálculo podem correr durante a noite, e o nosso objectivo (auto-proposto) é tornar o sistema interactivo em tempo real. Assim, todos os dados necessários a uma visualização célere e interactiva são pré-calculados. Médias, desvios-padrões, mínimos, máximos, e limites de outliers para ambas as distribuições e para todos os níveis de sensibilidade e granularidade. Ainda assim, para conjuntos de tipos de utilizador (há milhares de utilizadores "normais") e granularidades "finas" (144 períodos de 10 minutos por dia), pode levar vários minutos a abrir a página que lista os outliers. Foi desenvolvido um sistema de cache que guarda a página pré-gerada e se necessário é trivial adaptar para gerar à partida todos os relatórios possíveis fazendo um script que invoca o wget – no entanto, dada a escassez de espaço em disco foi considerado desaconselhável.

4.5 Implementação: Identificação de queries suspeitas

Diferentes tipos de utilizadores fazem diferentes tipos de queries. Os DBA e "power users" têm esses privilégios para fazer tarefas de manutenção e administração, pelo que muitas das suas queries são comandos especiais, ou ainda criação de tabelas, importação de dados, etc. Apenas foram registadas cerca de 70 queries do tipo SELECT em quatro meses, que tivessem efetivamente retornado dados. Não é suposto os DBA fazerem SELECT a dados sensíveis, pelo que este valor extremamente baixo é normal e desejável. Fazem outro tipo de queries que não são relevantes para este trabalho, como por exemplo as queries que apenas escrevem dados sem retornar nenhum para o utilizador, CREATE TABLE, INSERT, UPDATE, ou ainda comandos especiais de administração, etc.

Similarmente, diferentes bases de dados, com diferentes tabelas, requerem SELECTs diferentes. Os utilizadores aplicacionais fazem padrões bem definidos de SELECTs, embebidos na aplicação correspondente a esse utilizador. Os utilizadores "normais", teoricamente sem privilégios, apenas são registados pelo Guardium por haver objectos públicos, tabelas configuradas como sendo de acesso público.

As queries mais suspeitas seriam, por hipótese, algo introduzido manualmente no uso interativo de um cliente de acesso a bases de dados. Não é expectável, nem foi empiricamente verificado, que existam com frequência SELECTs a dados sensíveis feitos à mão. Os acessos a estas bases de dados são geralmente feitos por aplicações com um pequeno conjunto de padrões bem-conhecidos que ocorre milhões de vezes.

Partimos assim do pressuposto que uma query anómala, ou suspeita, é necessariamente um comportamento raro e provavelmente uma query feita à mão por conveniência ou para satisfazer uma curiosidade pessoal.

4.5.1 Importação e transformação de dados

É importado diariamente um ficheiro CSV chamado G_SQL, gerado pelo Guardium, com as queries feitas no dia anterior usando o comando mysqlimport. Por insuficiência de recursos de hardware no servidor dedicado ao projecto, o texto completo das queries não é retido durante mais de um dia, o que não causa problemas porque essa informação já está na base de dados central, situada no servidor Pulso.

A transformação das queries em padrões é feita no Pulso com os seguintes critérios, que evitam falsos positivos ao eliminarem variações que em nada afetam o comportamento da query:

1. Substituir algarismos de 0 a 9 por "_" (*underscore*)
2. Converter o texto da query para letra maiúscula
3. Remover espaço em branco (*whitespace*)
4. Calcular o CRC32

4.5.2 Classificação de queries

O objetivo primário deste trabalho é identificar acessos a dados sensíveis por parte de utilizadores privilegiados, como DBAs, que exercem a sua atividade sobre eles

diretamente na consola, introduzindo comandos manualmente. Focámo-nos então sobre queries raras, ajustando critérios em conjunto com a equipa de coordenação para submeter a inspeção manual apenas uma quantidade de queries que fosse exequível examinar.

Para classificar as queries por ordem de gravidade, foram então definidas as seguintes heurísticas:

- Comprimento da query menor do que 3 KB. Os padrões de queries provenientes de aplicações são fáceis de identificar visualmente por um operador do sistema que já esteja habituados a vê-las todos os dias. Todas as queries suspeitas verificadas foram inferiores a 1KB, pelo que 3 KB já é mais do que o suficiente neste contexto.
- Frequência de ocorrência. Foram desenvolvidas duas classificações, diária, e global. No primeiro dos interfaces apenas se exibem queries que ocorram até 3 vezes no mesmo dia, inclusive. No segundo, mostram-se as que ocorreram menos de 100 vezes desde o início da monitorização, mais uma vez para ter margem de manobra.

De seguida examinaremos como resultado final deste trabalho, as estatísticas que dizem respeito ao número de anomalias encontrado, agregado por mês e por base de dados, para comparar a evolução da situação de cada base de dados ao longo do tempo.

Capítulo 5 Resultados

Serão apresentados neste capítulo comparações de agregados mensais para quatro métricas que quantificam alguns aspetos da segurança das bases de dados da PT Comunicações sob monitorização da plataforma IBM Guardium no âmbito do projeto DAMS da DCY.

Estas métricas são idênticas ás que geram os "mapas de calor" alojados no sistema Pulso, que não seriam adequados para apresentar neste relatório dado a sua extensão e o seu caráter efémero.

Os eixos verticais dos gráficos estão em escala logarítmica; indicam a ordem de grandeza da métrica, ou seja, a gravidade que um conjunto mensal de ocorrências possa ter.

O primeiro mês de cada gráfico poderá apresentar um valor inferior aos outros, em virtude da monitorização não ter iniciado no primeiro dia do mês.

Estes gráficos condensam uma enorme quantidade de dados que seria impraticável apresentar – milhares de anomalias. Os algoritmos foram afinados ao longo dos vários meses e aplicados com efeito retroativo a todo o conjunto de dados à medida que foi possível excluir mais falsas positivas. Assim, a evolução dos dados apresentados não foi afetada pelo processo de desenvolvimento, apesar de recolhida ao longo dos vários meses em que este decorreu.

5.1 Extrações massivas - Row Outliers

Os Row Outliers são consultas que retornam um volume de dados que seja muito superior ao normal (três desvios-padrões) e possa indicar uma fuga de dados.

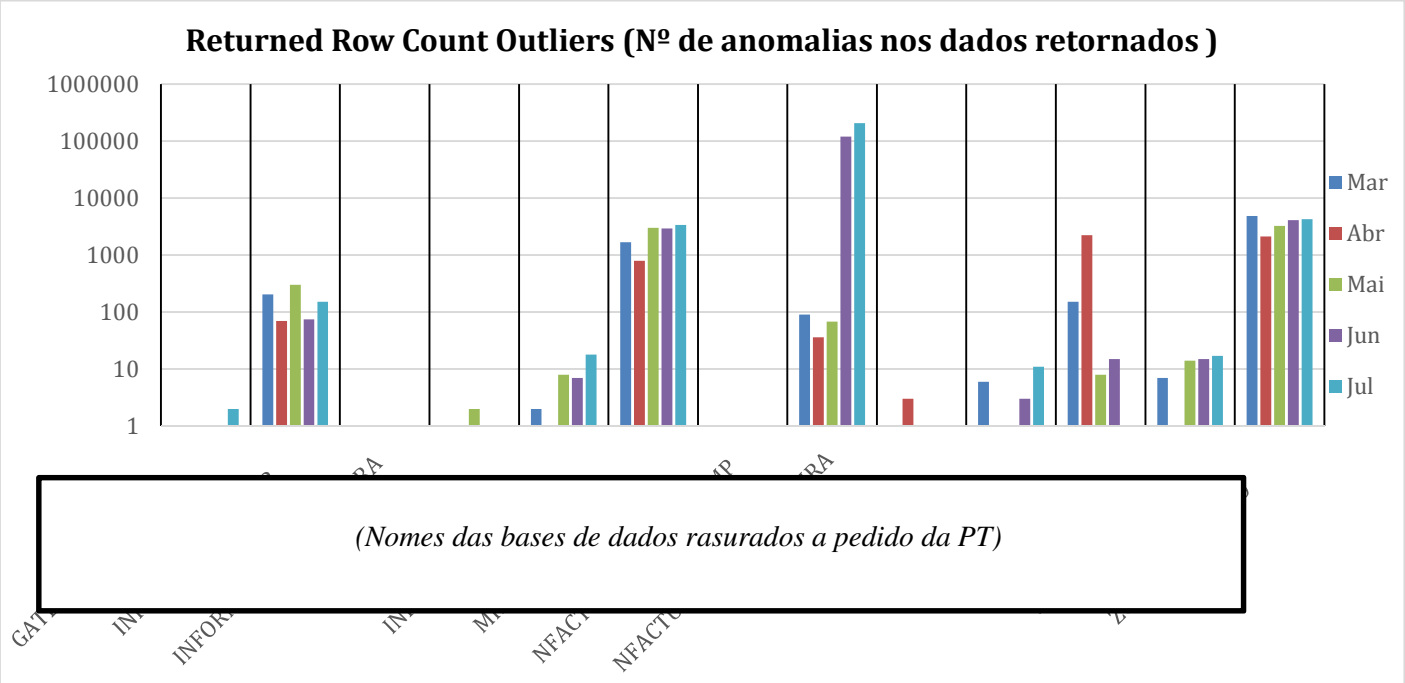

Figura 7 – *Nº de extrações massivas de dados entre Março e Julho, por base de dados.*

Na figura 7 podemos observar que apenas 13 das 17 bases de dados sob monitorização têm consultas, por vezes, com resultados muito superiores á média. Por usarmos uma escala logarítmica, bases de dados só com uma anomalia aparecem vazias, visto que o logaritmo de um é zero.

Apesar da ordem de grandeza destas ocorrências ainda ser apreciável, ronda apenas 1% do total das consultas a dados sensíveis, previamente expostas na figura 2.

No entanto, a quantidade expectável de anomalias com este método seria no máximo 0.3%. Existem muitos falsos positivos, quer pela organização dos sistemas monitorizados, quer pela qualidade dos dados obtidos:

- 63,2% dos utilizadores não acedem o suficiente para serem caracterizados com precisão.
- Há sempre novos utilizadores a aparecer, e leva tempo a recolher dados suficientes.
- Muitos acessos são feitos via aplicações, o que oculta o comportamento dos utilizadores.

- É legítimo haver consultas ao extrato mensal superiores à média (p. ex. clientes empresariais).
- Pode haver consultas que peçam um extrato com vários meses.
- Também é normal serem feitas cópias de segurança (backups).
- Novas versões de um sistema podem introduzir alterações comportamentais

No entanto, apesar das falsas positivas, uma extração verdadeiramente massiva sobressai imediatamente. As anomalias são exibidas diariamente por ordem de gravidade, da maior para a menor. As mais distantes do normal, presumivelmente maiores são exibidas primeiro. E mesmo que não sejam detetadas anomalias, para nos precavermos contra uma eventual falha do algoritmo, o sistema exibe na mesma as 20 extrações mais massivas do dia.

5.2 Queries Raras – Rare Queries

As Rare Queries são consultas a uma base de dados compostas por um padrão estrutural de comandos que raramente ocorre e pode ser um comportamento suspeito. Esse padrão é obtido filtrando certas variações que não afetam o significado das instruções que compõem a consulta, por exemplo, questões de formatação do texto, ou parâmetros numéricos como o número de telefone, a data, etc.

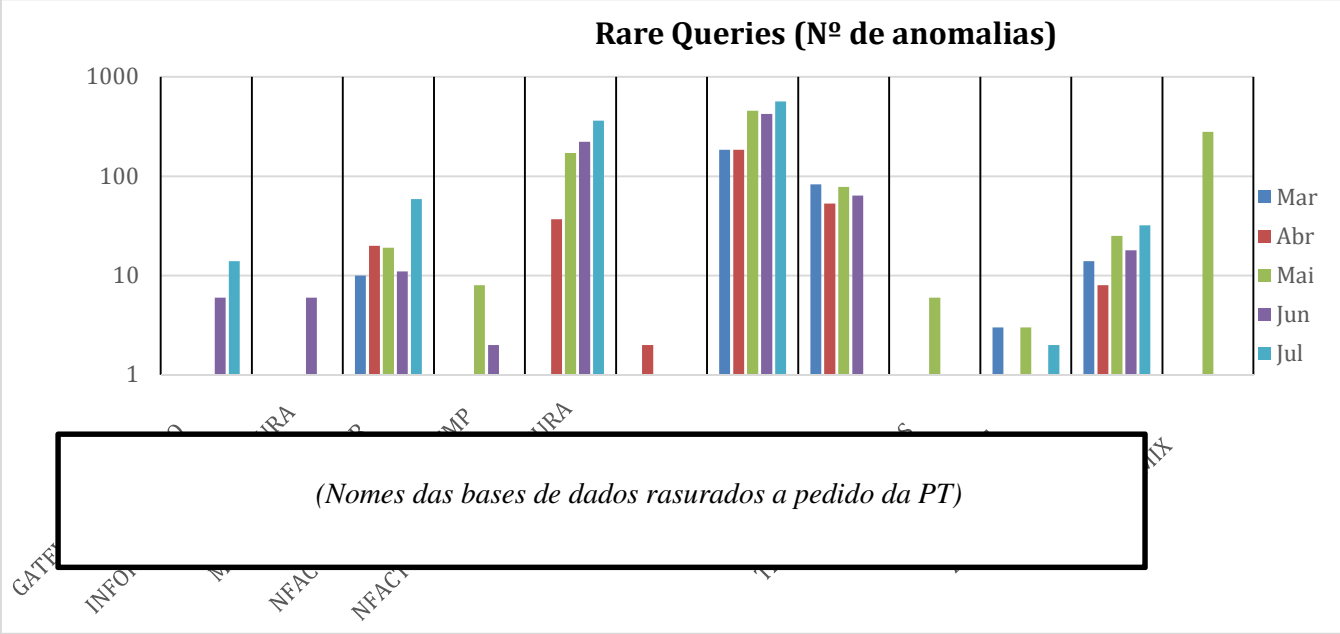

Figura 8 – *Nº de padrões raros de instruções de consulta a dados sensíveis entre Março e Julho.*

A Figura 8 mostra que apenas 12 das 17 bases de dados registam anomalias segundo este critério – queries cujo padrão estrutural de comandos, livre de parâmetros numéricos, não é usual numa determinada base de dados. Foram selecionadas como anómalas e sujeitas a inspeção manual as consultas definidas por:

1. Conjuntos de instruções de comprimento inferior a 3000 caracteres.
2. Frequência do padrão igual ou inferior a 3 vezes por dia.

Estes critérios foram definidos em conjunto com a equipa de orientação para tentar identificar apenas instruções suficientemente pequenas para terem sido introduzidas manualmente no sistema por um utilizador potencialmente mal-intencionado, e também para possibilitar a sua rápida inspeção manual.

Note-se que este número já dá uma margem de manobra: foram feitas experiências ao longo dos meses, e apenas 1000 caracteres já seriam suficientes para conter as instruções que parecem suspeitas.

Ao contrário do método anterior, que se foca na extração massiva de dados, este permite:

- Identificar extrações não-massivas que possam ainda assim ser ilegítimas.
- Tolerar grande variabilidade nos resultados obtidos pelo mesmo tipo de consulta.
- Garantir deteção fiável de anomalias mesmo para utilizadores com poucos acessos.
- Submeter a rápida supervisão manual 100 vezes menos anomalias.

Assim, pode-se concluir que este método é mais útil pois abrange a totalidade do comportamento e não apenas o volume de dados que retornou. Qualquer extração massiva seria, em princípio, também uma query rara. Por ex. um SELECT * feito por um utilizador que nunca consulta a tabela inteira.

De seguida examinaremos algum trabalho adicional desenvolvido neste projeto, ao longo das mesmas linhas de orientação que os objetivos principais: três métricas auxiliares que também estão a ser registadas no sistema Pulso.

Capítulo 6 Trabalho adicional

Como nota de rodapé, gostaríamos de apresentar algumas métricas cuja recolha foi também considerada útil pela PT.

6.1 Contagem de objetos públicos - Public Objects

Os Public Objects são campos e tabelas de uma base de dados, contendo dados sensíveis e configurados como acessíveis a qualquer utilizador da base de dados.

Na Figura 9 observa-se que os objetos públicos não têm sofrido alteração significativa em todas as bases de dados, desde Abril a Julho. No entanto, o número ideal seria zero. Um utilizador não-privilegiado não deveria ter acesso a dados sensíveis, vindos diretamente da base de dados, apenas deveria aceder através de aplicações.

O caso é tornado menos claro pelo facto de haverem aplicações que retransmitem o utilizador em vez de registarem os acessos em nome da aplicação.

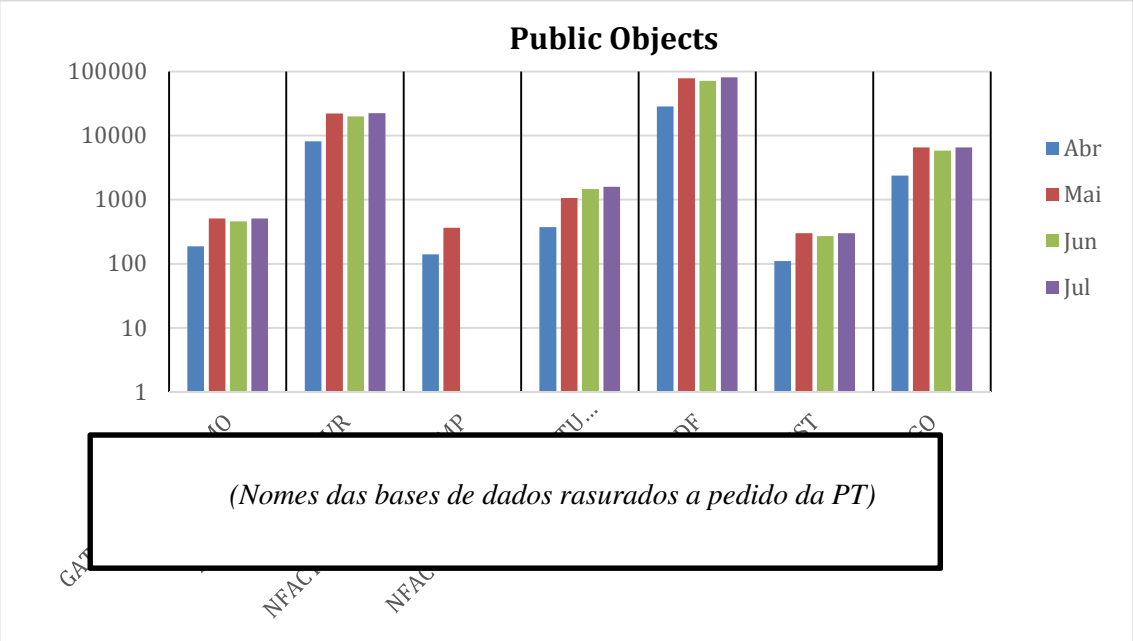

Figura 9 – *Nº de objetos públicos detetados em cada mês, entre Abril e Julho, para cada base de dados.*

6.2 Consultas a dados sensíveis - Sensitive Selects

Os Sensitive Selects são o número em bruto de consultas que retornam dados sensíveis, sejam ou não potencialmente suspeitas de acordo com qualquer um dos métodos utilizados.

São ignoradas, por acordo prévio com os orientadores, consultas que não retornem dados:

- Acessos do tipo SELECT que não retornam dados sensíveis.
- Instruções de outros tipos: INSERT, UPDATE, DELETE, etc.
- Instruções inválidas que retornam erro.

A variação depende apenas do tráfego da base de dados e não constitui qualquer anomalia: Sendo estes os dados em bruto, não está a ser aplicado qualquer método de deteção de anomalias para os filtrar.

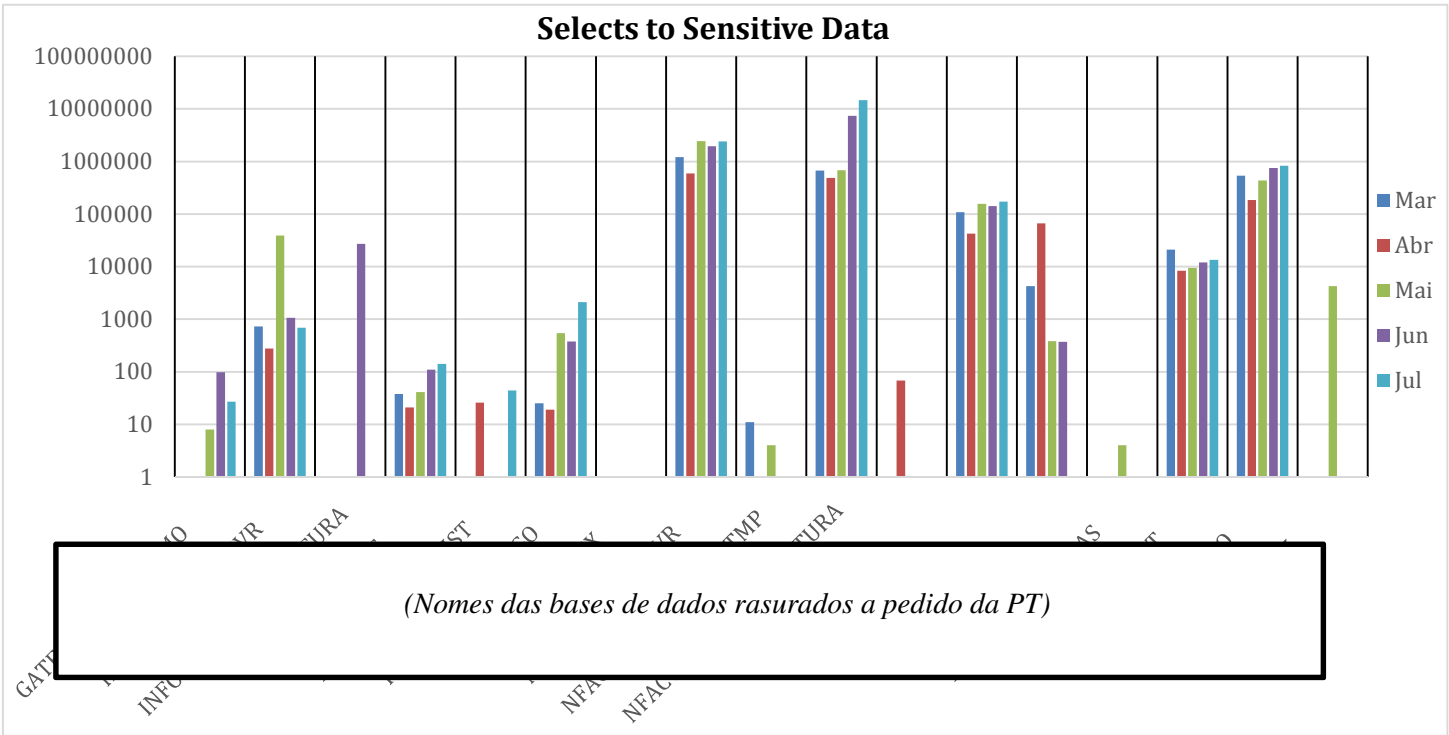

Figura 10 – *Nº de acessos a dados sensíveis entre Março e Julho, sem filtragem.*

Foi verificada uma alteração drástica num utilizador: em vez de extrair quantidades de dados maiores por cada consulta, passou a fazer regularmente um número muito superior de consultas, cada uma delas com extrações muito menores de dados. Isto pode ser observado na figura 10, base de dados NFAC(…)TURA, para Junho e Julho.

6.3 Métrica adicional – Misused Users

Foram contabilizadas as horas totais de utilização por utilizadores não-aplicacionais como heurística para deteção de aplicações não autorizadas, sem se registar nenhum comportamento especialmente anómalo. Apenas foi encontrado um utilizador não-aplicacional que excedeu as 40 horas de trabalho, registando 46 horas em apenas uma das muitas semanas de monitorização. Assim sendo, não há dados suficientes para apresentar um gráfico comparativo das bases de dados ao longo dos meses.

Capítulo 7 Conclusão

O projeto foi desenvolvido dentro do prazo e com resultados satisfatórios. Apesar de haver falsas positivas, as situações verdadeiramente anómalas são detetadas facilmente através da ordenação dos resultados diários que dá prioridade a casos mais sérios, como extrações muito massivas, ou queries muito raras. Segue uma breve apreciação crítica dos métodos utilizados, com uma sugestão para trabalho futuro.

7.1 Extrações massivas - Método dos períodos homólogos

Verificou-se experimentalmente através de testes com várias granularidades temporais que o método proposto não tem resultados melhores que uma simples média global, antes pelo contrário, o seu uso para deteção de extrações massivas pode ser uma escolha pior:

1. O método dos períodos homólogos não se aplica aos dados retornados. Vejamos porque é que não podemos assumir que um determinado volume de dados retornados implica que este tenha sido feito a uma determinada hora, e vice-versa:
   - O volume de uma extração não implica que haja uma extração feita à mesma hora em dias diferentes, mesmo dia em semanas diferentes, etc. Não implica que volumes semelhantes sejam feitos a horas ou dias semelhantes (homólogos). O nosso extrato mensal, por exemplo, tem exatamente as mesmas chamadas a qualquer hora de qualquer dia.
   - O contrário também é verdade, extrações feitas em períodos de tempo homólogos não retornam necessariamente o mesmo volume. Uma query pode consultar qualquer tipo de extrato de qualquer tipo de consumidor, que faça mais ou menos chamadas, ou que abranja vários meses, ou seja filtrado por vários critérios.
2. Dividir as amostras em períodos de tempo multiplica a quantidade de dados necessária para garantir alguma precisão na identificação de anomalias.
3. Um utilizador que faça apenas um acesso mesmo que seja uma extração massiva, não será dado como suspeito, pois a média de um único acesso é a própria média, o que não é estatisticamente anómalo. Para mitigar alguns destes casos, o programa mostra os 20 acessos mais massivos caso não haja anomalias a reportas.

7.2 Queries raras

A filtragem de queries pelo padrão opbitod depois de removidos os parâmetros numéricos, teve, no geral, bons resultados. Existe apenas uma aplicação (T…O) que gera queries computacionalmente. Neste caso, o padrão estrututal, mesmo depois de removidos os parâmetros, tem um elevado nível de variabilidade.

Note-se que o método desenvolvido apenas elimina parâmetros numéricos, pelo que tentar aplica-lo a outras situações pode não ter a mesma eficácia (queries que pesquisem em campos textuais, como nomes ou moradas). Idealmente a filtragem de parâmetros envolveria um interpretador de SQL, tarefa demasiado extensa para o âmbito deste projeto, e tornada complexa pela necessidade de suportar vários tipos de BD.

Capítulo 8 Trabalho futuro

Dada a elevada taxa de anomalias verificada pela métrica Row Outliers, foi proposta uma métrica composta - Rare Outliers - que englobasse ambos os critérios, tendo como objetivo eliminar falsas positivas na deteção de extrações massivas de dados. Na figura 11 podemos verificar que apenas 6 das 17 bases de dados sob monitorização registam extrações massivas correspondentes a padrões raros de instruções. Sendo que mais de 99,999% dos acessos são considerados normais por este critério conjunto, apresenta-se como uma solução fiável para a deteção de anomalias.

Uma inspeção manual dos resultados revela que estas anomalias partem usualmente de utilizadores aplicacionais. Mesmo os padrões estruturais das instruções podem variar bastante dentro de uma aplicação, se esta permitir muitas combinações diferentes de critérios de pesquisa.

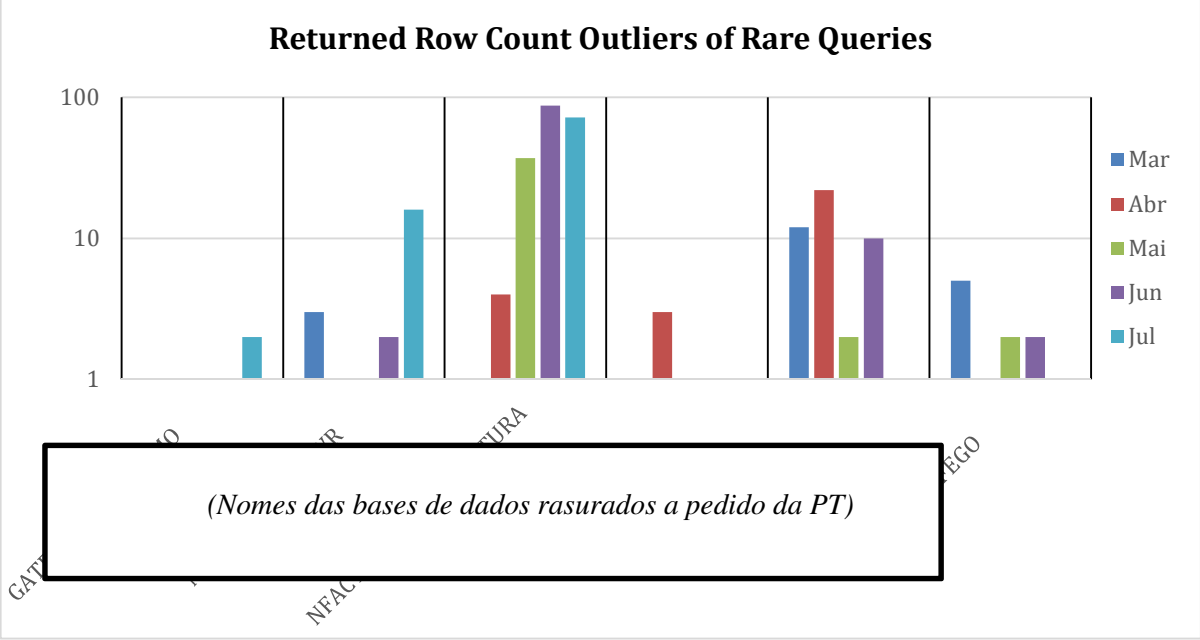

Figura 11 – *Nº de padrões raros de instruções que também extraem quantidades anómalas de dados.*

Dada a eficácia atingida pela composição destes dois critérios, e tendo em conta que o próprio Guardium usa algoritmos de *clustering*, seria vantajoso, num trabalho futuro, alargar o âmbito deste projeto para ter em conta outras dimensões do problema, por exemplo:

- Consultas feitas a partir de endereços IP pouco utilizados.
- Utilização de programas fora do comum para fazer essas consultas.
- Utilizadores que não costumam aceder a uma dada BD ou tabela.
- Frequência de acessos muito superior ao normal (extração massiva, por partes).
- Acessos feitos por um utilizador a horas fora do seu padrão de uso.

Bibliografia

[1] Ribeiro, Francisco G. T., Alegria, José, Correia, Miguel Pupo: *Descoberta e inferência de acessos anómalos a fontes de informação*, Tese de Mestrado, Faculdade de Ciências da Universidade de Lisboa, 2009

[2] Alegria, José A. S., Carvalho, Tiago F. R., Ramalho, Ricardo G.: *Uma experiência open source para "tomar o pulso" e "ter pulso" sobre a função sistemas de tecnologias de informação*, PT Comunicações, V Conferência da Associação Portuguesa de Sistemas de Informação (CAPSI), 2004

[3] Cathey, R., Ma, L., Goharian, N., Grossman D., *Misuse Detection for Information Retrieval Systems*, Illinois Institute of Technology, ACM 12th Conference on Information and Knowledge Management (CIKM), 2003

[4] Chung, C.Y., Gertz, M., Levitt, K., *DEMIDS: A Misuse Detection System for Database Systems*, University of California at Davis, 1999

[5] Helman, P., Liepins, G., Richards, W., *Foundations of Intrusion Detection*, Proceedings of the V IEEE Computer Security Foundations Workshop, 1992

[6] Čisar P. et al., *Skewness and Kurtosis in Function of Selection of Network Traffic Distribution*, Acta Polytechnica Hungarica Vol. 7, No. 2, 2010

[7] McAlister, D., *The law of geometric mean*, Proceedings of the Royal Society of London 29, 367-376, 1879

[8] Hodge, V.J., Austin, J., *A Survey of Outlier Detection Methodologies.*, Artificial Intelligence Review, Volume 22, Issue 2, pp. 85-126, 2004

[9] Oded S. et al, *Privileged yet Unauthorized - Outliers Mining for Database Activity*, IBM Research, 2013

[10] Ihara, S., *Information theory for continuous systems,* p. 2, World Scientific, 1993

[11] Jaynes, E. T., *Information Theory and Statistical Mechanics,* Physical Review, nº 106, pp. 620-630, 1957

[12] Gibrat, R., *Les Inégalités économiques*, 1931

[13] Lotti, F., Santarelli, E., Vivarelli, M., *Gibrat's Law in a medium-technology industry: empirical evidence for Italy*, "Entrepreneurship, Growth, and Innovation International Studies in Entrepreneurship", Volume 12, pp. 149-164, 2006

[14] Sutton, J., *Gibrat's Legacy*, Journal of Economic Literature, Vol. 35, pp. 40-59, 1997

[15] Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., *From Data Mining to Knowledge Discovery in Databases*, 1996

[16] Hastie, T., Tibshrani, R., Friedman, J., *The Elements of Statistical Learning*, p. xi, Springer, 2008

[17] Luhn, H.P., *A Business Intelligence System*, IBM Journal, 1958

[18] Rud, O. et al, *Business Intelligence Success Factors: Tools for Aligning Your Business in the Global Economy*, Wiley & Sons, 2009

[19]Watson, H.J., Wixom, Barbara H., *The Current State of Business Intelligence*, IEEE Computer, 2007

[20] Cates, J. E. et al, *The Ladder of Business Intelligence (LOBI): a framework for enterprise IT planning and architecture*, Int. J. Business Information Systems, 2005